



University of Warwick institutional repository: <http://go.warwick.ac.uk/wrap>

**A Thesis Submitted for the Degree of PhD at the University of Warwick**

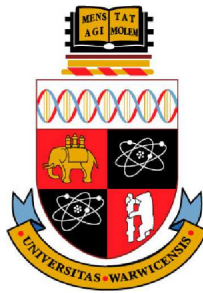
<http://go.warwick.ac.uk/wrap/1066>

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it. Our policy information is available from the repository home page.

# Statistical Inference from Large-Scale Genomic Data



Yinyin Yuan

Department of Computer Science

University of Warwick

A thesis submitted for the degree of

*Doctor of Philosophy*

March, 2009

# Contents

<b>Abbreviations</b>	<b>xii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Genome Architecture and Functions . . . . .	4
1.1.1 Key Processes and Related Data . . . . .	5
1.1.2 Microarray Data Acquisition . . . . .	9
1.2 Statistical Inference for Functional Genomics . . . . .	10
1.2.1 Tight Clustering of Gene Expression Profiles . . . . .	11
1.2.2 Clustering Validation Using Functional Annotation . . . . .	12
1.2.3 Transcriptional Regulatory Network Reconstruction . . . . .	13
1.3 Thesis Overview . . . . .	16
1.3.1 Thesis Contributions . . . . .	16
1.3.2 Thesis Organisation . . . . .	18
<b>2 Partial Mixture Model for Tight Clustering Gene Expression</b>	<b>19</b>
2.1 Introduction . . . . .	19
2.2 Existing Methods and Future Needs . . . . .	21
2.2.1 Linear Models . . . . .	21

2.2.1.1	Spline model . . . . .	22
2.2.1.2	Autoregressive model . . . . .	23
2.2.2	Parameter Estimation . . . . .	24
2.2.3	Limitations of Existing Methods . . . . .	26
2.2.4	Emergence of Tight Clustering . . . . .	27
2.3	Proposed Tight Clustering Method . . . . .	29
2.3.1	Minimum Distance Estimator (MDE) . . . . .	30
2.3.2	Weighted Mixture Model with MDE . . . . .	31
2.3.3	Partial Mixture Model with MDE (PMDE) . . . . .	34
2.3.3.1	The spline regression model . . . . .	34
2.3.3.2	The stopping criteria . . . . .	36
2.3.4	Experimental Validation of MDE with partial modelling . . . . .	37
2.3.5	The PMDE Clustering Algorithm . . . . .	41
2.4	Experimental Results . . . . .	42
2.4.1	Experiment on Simulated Data . . . . .	42
2.4.2	Experiments on Yeast Cell Cycle (Y5) Data Set . . . . .	44
2.4.2.1	Clustering yeast Y5 data set . . . . .	46
2.4.2.2	Gene ontology enrichment analysis . . . . .	50
2.4.2.3	Predictive accuracy test . . . . .	54
2.4.2.4	Scattered genes . . . . .	59
2.4.2.5	Comparative evaluation on scattered gene detection . . . . .	62
2.4.3	Experiments on Yeast Galactose Data . . . . .	63
2.5	Conclusions . . . . .	68
<b>3</b>	<b>Quantitative Assessment of Clustering Based on Gene Ontology</b>	<b>70</b>

3.1	Introduction . . . . .	70
3.1.1	An Introduction to Gene Ontology (GO) . . . . .	72
3.1.2	Rationales for GO-based Clustering Validation . . . . .	74
3.2	Existing Methods Assuming GO as Functional Categories . . . . .	76
3.3	Existing Methods Based on GO Semantic Similarity . . . . .	79
3.3.1	GO Semantic Similarity . . . . .	79
3.3.2	Problems of Methods in this Category . . . . .	81
3.3.3	Experimental Assessment . . . . .	83
3.3.3.1	Clustering validation indices . . . . .	83
3.3.3.2	Experiment . . . . .	84
3.4	Proposed Validation Method . . . . .	87
3.4.1	GO-based Term-Term Distance . . . . .	89
3.4.2	Within-Cluster Compactness . . . . .	91
3.4.3	Between-Cluster Similarity . . . . .	92
3.4.4	Combined Index WB . . . . .	93
3.4.5	Confidence Thresholds . . . . .	94
3.5	Experimental Results . . . . .	95
3.5.1	Evaluation of Six Clustering Algorithms . . . . .	97
3.5.1.1	Experiments on yeast Y5 data . . . . .	99
3.5.1.2	Experiments on Arabidopsis diurnal data . . . . .	103
3.5.2	Perturbation Experiment . . . . .	109
3.5.3	Finding Optimum Number of Clusters . . . . .	111
3.6	Conclusions . . . . .	114

## **4 A Bayes Random Fields Approach for Integrative Large-Scale Regulatory**

<b>Network Reconstruction</b>	<b>116</b>
4.1 Introduction . . . . .	116
4.2 Data Sources and Existing Methods . . . . .	119
4.2.1 Heterogenous Data Sources . . . . .	119
4.2.2 Existing Methods for Network Reconstruction . . . . .	120
4.2.2.1 Methods for single data source . . . . .	121
4.2.2.2 Methods for multiple data sources . . . . .	122
4.2.3 Existing Problems and Prelude to the Proposed Approach . . . . .	124
4.3 Proposed Bayes Random Fields (BRFs) Integrative Method . . . . .	125
4.3.1 Bayes Framework . . . . .	126
4.3.2 Random Fields Model . . . . .	128
4.3.3 A Gibbs Sampling Algorithm for BRFs . . . . .	129
4.4 Experiments . . . . .	133
4.4.1 Synthetic Networks . . . . .	134
4.4.2 <i>Saccharomyces Cerevisiae</i> Regulatory Network . . . . .	136
4.5 Conclusions . . . . .	146
<b>5 Conclusions and Future Research</b>	<b>147</b>
5.1 Conclusions . . . . .	147
5.1.1 Partial Mixture Model Tight Clustering . . . . .	148
5.1.2 Clustering Validation Using Gene Ontology . . . . .	149
5.1.3 Transcriptional Regulatory Network Reconstruction . . . . .	151
5.2 Future Research . . . . .	152
5.2.1 Inferring Causal Relations from Large-scale Gene Expression Data . . . . .	153

5.2.2	GO-driven Validity Index for Regulatory Network Inference	
	Methods . . . . .	154
5.2.3	Combined Analysis of DNA Sequence and Microarray Data .	154
<b>A</b>		<b>157</b>
A.1	Theoretical Comparison between MDE and MLE . . . . .	157
A.2	Mean Integrated Squared Error . . . . .	158
<b>B</b>		<b>161</b>
B.1	Efficient Computation of Partial Correlation . . . . .	161
<b>References</b>		<b>186</b>

# List of Figures

1.1	Key processes in the central pathway (adapted from [8]). . . . .	4
1.2	Genomic data is providing large-scale descriptions of nearly all components and interactions within the cell (adapted from [75]). . . . .	6
1.3	An Arabidopsis circadian gene network of six genes (adapted from Locke <i>et al</i> [87]). Circles are proteins and DNA segments represent genes. . . . .	14
2.1	Comparing MDE and MLE by data fitting and their residual histograms. (a) MDE fit (pink line) and MLE fit (blue line) to simulated data generated from three sine waves; (b) Histogram of residuals by MDE; (c) Histogram of residuals by MLE; (d) MDE fit (pink line) and MLE (blue line) fit to simulated data generated from two sine waves; (e) MDE fit (pink line) and MLE (blue line) fit to data with many outliers; (f) MDE fit (pink line) and MLE (blue line) fit when two components are of same size. . . . .	40



2.2	The resulting partition by the partial regression clustering algorithm for the simulated data set. The first 6 plots correspond to the gene clusters, the left plot in the third row shows the outliers, the right plot in the last row shows the whole data set. . . . .	43
2.3	The original partition of the yeast Y5 data set with the bottom right plot of the whole data set. . . . .	46
2.4	The clusters by the partial regression clustering algorithm for the Y5 data set. The bottom right plot shows the scattered genes. . . . .	47
2.5	Heatmaps for the original partition (left), SplineCluster (middle) partition and the proposed PMDE clustering (right) partition. The brighter red color corresponds to higher expression levels and brighter green color corresponds to lower expression levels. . . . .	48
2.6	Predictive accuracy plots for five clustering methods on Y5 data set. Five clustering methods are evaluated in terms of their functional group prediction accuracy. The five methods are the proposed PMDE (red), SplineCluster (violet), MCLUST (black), hierarchical clustering (green), and K-means (blue). The higher the curve is the better the performance is. . . . .	55
2.7	The profiles of seven genes related to Late G1, SCB regulated cell cycle phase. The red profile is the gene “TIP1/YBR067C”, one of the ten scattered genes. It displays a distinctive pattern from the other six genes annotated to be in the same functional group. . . . .	59

2.8	Comparison of performance of PMDE and MCLUST in outlier detection. A small index value of $WS S$ indicates better performance in outlier filtering. PMDE performs better than MCLUST with large number of clusters. . . . .	62
2.9	Expression data across 20 time points in four functional categories of yeast galactose data. . . . .	64
2.10	Scattered genes in original cluster 2 of the yeast Galactose data set. The expression profiles of some scattered genes detected by the proposed algorithm are plotted for the yeast Galactose data set. This plot shows the expression patterns of all 15 genes in original cluster 2, among them the 3 colored genes are the detected scattered genes. . .	65
2.11	Scattered genes in original cluster 3 of the yeast Galactose data set. The expression profiles of the 3 scattered genes in original cluster 3. They share GO annotations but have various expression patterns. . . .	66
3.1	The graph structure of GO, edge weights are to be defined in Section 3.4.1 . . . . .	73
3.2	An example of functional overlapping in gene clusters with the over-represented terms (pink) for three gene clusters (C1, C2 and C3). There is an overlapping over-represented term (GO:0000278) between C1 and C2. . . . .	75

3.3	Experiments on discriminating random partitions (yellow curves) from meaningful partitions (non-yellow curves) with semantic similarity based on the Silhouette index. For each of the GO category, three semantic similarity measures, Resnik's (R), Lin's (L), Jiang and Conrath's (JC) measure, are used. . . . .	85
3.4	Experiments on discriminating random partitions (yellow curves) from meaningful partitions (non-yellow curves) with semantic similarity based on the Davies-Bouldin index. Colour codes are provided in the legend in Figure 3.3. . . . .	86
3.5	Experiments on discriminating random partitions (yellow curves) from meaningful partitions (non-yellow curves) with semantic similarity based on the Dunn index. Colour codes are provided in the legend in Figure 3.3. . . . .	87
3.6	The <i>Arabidopsis L. Heyn</i> th diurnal data clustered into eight clusters by K-means clustering. . . . .	96
3.7	For the Yeast Y5 data set, plots of (a),(b),(c) WCC scores and (d), (e), (f) BCS scores for six clustering algorithms and the average of ten random runs based on the three GO categories BP, MF and CC, respectively. . . . .	102
3.8	For the yeast Y5 data set, normalised scores of six validity indices for various clustering algorithms and random partitions. The solid lines denote that the indices are GO-driven, while the dashed lines denote data-driven indices. . . . .	103

3.9	For the Arabidopsis diurnal data set, normalised scores of six validity indices for various clustering algorithms. Colour codes indicating the validity index identities are the same as they are in Figure 3.8. . . . .	105
3.10	Normalised scores of validity indices with increasing level of perturbation in the yeast galatose data set. Large values correspond to good partitions for all the indices. (a) GO-driven validity indices WB, BHI and BSI calculated based on the GO category ‘Biological process’, (b) WB index and data-driven indices. . . . .	110
3.11	Scores of the six validity indices as a function of cluster numbers for the yeast Y5 data set using (a) SplineCluster algorithm, (b) hierarchical algorithm, (c) PMDE algorithm. Colour codes are the same as they are in Figure 3.8 (black: WB, red: BHI, green: BSI, dark blue: Dunn, light blue: CH, pink: DB). . . . .	113
4.1	Experimental results for the synthetic networks. . . . .	135
4.2	Experimental results for the 909 yeast genes. . . . .	139
4.3	Plot of the inferred cell cycle specific sub-network, with size of the nodes indicating the degree of connectivity. . . . .	140
4.4	ROC curves for threshold selection methods for the three data types used in network reconstruction and the resultant network by BRFs on the 296-gene sub-network. PCOR stands for the partial correlation. . .	141
4.5	Connectivity degree distribution of the 296-gene sub-network. The $x$ axis shows the degree of connectivity of 296 nodes on $\log_2$ scale. . . .	142
4.6	Plot of time series data of the eight transcription factors (pink) and the genes they are regulating (grey). . . . .	144

# List of Tables

2.1	Cross tabulation of the original partition and the PMDE clustering partition for the Y5 data set. . . . .	46
2.2	Over-represented GO terms by the proposed PMDE algorithm for the Y5 data set . . . . .	52
2.3	Over-represented GO terms by the SplineCluster Algorithm for the Y5 data set . . . . .	53
2.4	Verified cell cycle related (68) genes in the yeast Y5 data set . . . . .	57
2.5	Cross-tabulation of clustering outcome (C1-C8 and SG) with verified gene functional categories for the yeast Y5 data set . . . . .	58
2.6	Details of the set of scattered genes for the yeast Y5 data set detected by PMDE, including their SGD IDs, the frequencies that they are found across eight experiments of various thresholds, and their annotations. .	61
2.7	Cross-tabulation of the original partition (O1-O4) and the resulting partition (C1-C4 and SG) for the yeast Galactose data set. . . . .	66
2.8	Over-represented terms in each original cluster for the yeast galactose data set. . . . .	67

3.1	Confidence levels ( $\alpha$ ) and corresponding p-value cut-offs ( $\rho$ ) in the PMDE partition for the Y5 data set. . . . .	99
3.2	Confidence levels ( $\alpha$ ) and corresponding p-value cut-offs ( $\rho$ ) in the PMDE partition for the Arabidopsis data set . . . . .	104
3.3	Over-represented GO terms in the K-means partition for the Arabidopsis data set . . . . .	107
3.4	Over-represented GO terms in the PMDE partition for the Arabidopsis data set . . . . .	108
3.5	Confidence levels ( $\alpha$ ) and corresponding p-value cut-offs ( $\rho$ ) in the starting partition for the Galatose data set for the perturbation experiment	109
4.1	Some parameter settings for SynTReN software to generate the simulated data sets, the rest are set as default. . . . .	134
4.2	Over-represented GO terms in the transcriptional modules for eight transcription factors. . . . .	143
4.3	Over-represented GO terms in four phases specific modules found in Figure 4.2(b). . . . .	145

## Acknowledgements

I would like to thank my supervisor, Dr Chang-Tsun Li, for giving me the opportunity to explore the vast spectrum of knowledge and discover the exciting field of bioinformatics. Since our acquaintance in 2003 during my master course, he has always been supportive and encouraging, not only for my Ph.D. study but also in pursuing my academic goals. I received his vision, experience, and in particular his rigorous research attitude, which will be greatly beneficial to my future career.

I am grateful to Prof Roland Wilson, who, as my advisor, helped me keep track of my progress in the last three years and supported me through many difficulties. I want to acknowledge the generous support and advices given by Dr Vicky Buchanan-Wollaston and Linda Hughes in Warwick Horticulture Research International. During our interactions, many ideas were inspired that have led to significant progress in my study. I am also thankful to Dr Sascha Otts in the Centre of Systems Biology whose penetrating insights in both computational and biological aspects of bioinformatics are very helpful.

I also would like to thank many colleagues and friends for their help and friendships, without which my life at Warwick University would not have been so fulfilling and memorable.

I dedicate this thesis to my parents for their unconditional love, support,  
and encouragement.



## Abstract

This thesis explores the potential of statistical inference methodologies in their applications in functional genomics. In essence, it summarises algorithmic findings in this field, providing step-by-step analytical methodologies for deciphering biological knowledge from large-scale genomic data, mainly microarray gene expression time series.

This thesis covers a range of topics in the investigation of complex multivariate genomic data. One focus involves using clustering as a method of inference and another is cluster validation to extract meaningful biological information from the data. Information gained from the application of these various techniques can then be used conjointly in the elucidation of gene regulatory networks, the ultimate goal of this type of analysis. First, a new tight clustering method for gene expression data is proposed to obtain tighter and potentially more informative gene clusters. Next, to fully utilise biological knowledge in clustering validation, a validity index is defined based on one of the most important ontologies within the Bioinformatics community, Gene Ontology. The method bridges a gap in current literature, in the sense that it takes into account not only the variations of Gene Ontology categories in biological specificities and their significance to the gene clusters, but also the complex structure of the Gene Ontology. Finally, Bayesian probability is applied to making inference from heterogeneous genomic data, integrated with previous efforts in this thesis, for the aim of large-scale gene network inference. The proposed system comes with a stochastic process to achieve robustness to noise, yet remains efficient enough for large-scale analysis.

Ultimately, the solutions presented in this thesis serve as building blocks of an intelligent system for interpreting large-scale genomic data and understanding the functional organisation of the genome.

## Statement of Originality

I hereby certify that all of the work described within this thesis is the original work of the author. Any published (or unpublished) ideas and/or techniques from the work of others are fully acknowledged in accordance with the standard referencing practices.

A handwritten signature in black ink, appearing to read 'Yinyin Yuan', with a stylized, flowing script.

Yinyin Yuan

March, 2009

## Publications

- Y. Yuan and C.-T. Li, “A Bayes Random Field Approach for Integrative Large-Scale Regulatory Network Analysis,” *Journal of Integrative Bioinformatics*, 5(2):99, 2008.
- Y. Yuan, C.-T. Li and R. Wilson, “Partial Mixture Model for Gene Expression Tight Clustering,” *BMC Bioinformatics*, 9:287, 2008.
- Y. Yuan and C.-T. Li, “Understanding Gene Clusters: An Investigation into Quantitative Assessment,” submitted to *Bioinformatics*, 2008.
- Y. Yuan and C.-T. Li, “Probabilistic Framework for Gene Expression Clustering Validation Based on Gene Ontology and Graph Theory,” in *Proc. of International Conference of Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 625-628, Las Vegas, US, 2008.
- Y. Yuan and C.-T. Li, “Partial Mixture Model for Tight Clustering in Exploratory Gene Expression Analysis,” in *Proc. of International Symposium on BioInformatics and BioEngineering (BIBE)*, 1061-1065, Boston, US, 2007.

## Abbreviations

BCS	Between-Cluster Similarity
BHI	Biological Homogeneity Index
BP	Biological Process
BSI	Biological Stability Index
CC	Cellular Component
CH	Calinski and Harabasz index
DNA	Deoxyribonucleic Acid
EM	Expectation-Maximisation
FC	Functional Compactness
FDR	False Discovery Rate
FS	Functional Similarity
GO	Gene Ontology
IC	Information Content
MDE	Minimum Distance Estimator
MF	Molecular Function
MLE	Maximum Likelihood Estimator
mRNA	messenger Ribonucleic Acid
PAM	Partitioning Around Medoids
PMDE	Partial mixture model with MDE
WCC	Within-Cluster Compactness

# Chapter 1

## Introduction

The genomics age has entered a new era to provide a grand picture of the whole genomes. Advances in microarray, deoxyribonucleic acid (DNA) sequencing techniques, and other high-throughput biotechnologies have brought great success to the life sciences. With the support of these high-throughput biotechnologies, significant breakthroughs in life science have been achieved, such as the advancement of cell reprogramming [139] and the development of low cost sequencing techniques [42]. Increasingly, high-throughput technologies are changing the biological landscape with their efficiency, cost effective nature and genome-wide coverage.

Therefore, some of the most significant advances in genomics research in recent years have been achieved with the availability of these high-throughput technologies to produce large-scale genomic data. The advent of these genome-scale data sources has transformed conventional biological research into data-oriented investigations. In these investigations, a key research direction is the effective interpretation and efficient utilisation of these information-rich data. For instance, inference from these data at the molecular level has been revolutionary in medicine [40], both because of the highly

---

informative nature and the comprehensive genome-wide coverage of the data.

Consequently, statistical inference from large-scale genomic data has emerged as a new discipline employing innovative data mining methods supported by high-throughput biological experimental technologies. The central goal is to employ computational techniques for extracting knowledge from the large-scale genomic data, and translate gained knowledge into system-based applications such as disease classification.

Statistical inference methods have been intensively applied to various research areas such as multimedia processing [150] and computational neuroscience [58]. Although statistics has been the support for biological data analysis for many years, biological data has changed over time not only in size, but above all in structure. In particular, genomic data from high-throughput biotechnologies have their unique, diverse features. New statistical challenges arise from the requirements of analysing these high-throughput genomic data, and, ultimately, deriving fundamental biological information. In this sense, innovative, objective and effective computational methods are urgently needed.

In recognition of this, this thesis addresses existing problems in statistical inference from large-scale genomic data resulting from high-throughput technologies, which, in essence, originate from the scale and the intrinsic characteristics of the data. In response, it introduces new computational statistical approaches built upon up-to-date biological understanding. The new algorithms have been developed taking into account the unique characteristics of genomic data, and have been validated by means of statistical benchmarking with both synthetic and real-world data. Above all, the thesis represents the methodology of designing novel statistical models in accordance with biological prior knowledge about the subjects under investigation.

The main theme of this thesis is the application of statistical methodologies to ge-

nomics research, bridging multiple disciplines such as computer science, molecular biology and statistics. The thesis highlights an exposition that advanced statistical and computational techniques, combined with highly problem-specific modeling efforts, can be eventually developed into elegant yet realistic formulations for genomics research. On the other hand, the immense complexity and stochasticity nature of the data faced by genomics research not only challenge the fields of theoretical and algorithmic statistical learning, but also foster new developments within. Further, insights gained from this process could lead to new perspectives in algorithmic findings for a broad range of fields that involve statistical learning, such as signal processing and neuroscience.

This chapter is organised as follows. To understand the central goal in genomics, this chapter first gives a brief description about the genome architecture and functionality. It then discusses the current state as well as the strengths and weaknesses of the representative data types presented, with an emphasis on how the genomic data are related to certain biological process and how they contribute to the analysis. Specifically, this chapter gives a description about the acquisition process of gene expression data from microarray technology, the main source of data analysed in this thesis, to help understand the data particular characteristics. Later the key issues related to genomic data analysis are laid out, as both derived from the literature and initial experimental data analysis. Following this, the objectives of this thesis are presented. The chapter concludes with a thesis outline and a description of chapter connections.

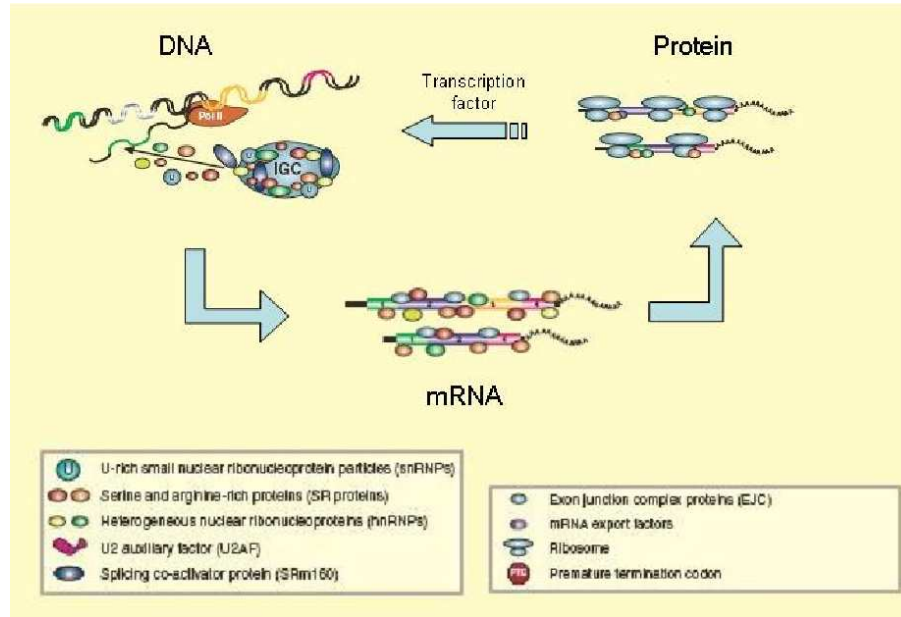


Figure 1.1: Key processes in the central pathway (adapted from [8]).

## 1.1 Genome Architecture and Functions

Functional genomics is, fundamentally, an area of research dedicated to understanding the structure and functional organisation of the genome [40]. The central dogma states that it is the genetic information encoded in the genes, which through a molecular decoding process, facilitates the functioning of cells in a living organism [27]. Deciphering the gene control circuitry encoded in the genes and its functional organisation is a fundamental problem in genomics research, and is a focus of this thesis.

The expression of genetic information encoded in the genes occurs in two stages, as depicted in Figure 1.1. Genes are segments of DNA, which is a long double-stranded anti-parallel molecule in which single complementary strands reversibly bind to each other to form a double stranded helix. From the left of Figure 1.1, genes are regulated by their own gene regulatory proteins, namely transcription factors, and transcribed into messenger ribonucleic acid (mRNA). This is the transcription stage, which refers



to the process of making a single-stranded mRNA molecule using a single coding DNA strand as a template, it is also the initial step of gene expression. mRNA are then translated into proteins which are responsible for carrying out nearly all cell functions. In turn, some of the proteins can again act as transcription factors which act to (coordinately) regulate transcription itself. These transcription factors regulate the next gene expression for the gene themselves are encoded by and/or the transcription of other genes. The whole procedure is governed by complex biochemical interactions that regulate gene expression and interaction. Therefore, the regulatory mechanisms are vital in directing genetic information flow and are the key to a global understanding of genome functions.

Study of the above genetic information flow from gene to protein in the central pathway helps reveal functional regulatory components in the genome, discover their connections with each other, and ultimately lead to mapping out the whole picture of the regulatory mechanism. The fundamental problem is how to infer collective gene regulatory functions and clarifying the roles of genes in cellular processes. By providing computational methodologies applicable to the high-throughput data that monitor these processes, this thesis aims to shed light on the study of gene regulatory mechanisms. The investigation focus is on transcriptional activities that are central to the regulatory mechanisms in the genome.

### 1.1.1 Key Processes and Related Data

Recent advances in high-throughput technologies have enabled the entire information flow procedure in the central pathway, as described in Section 1.1, to be captured on a genome-wide scale. To map out how the data is monitoring different cellular processes

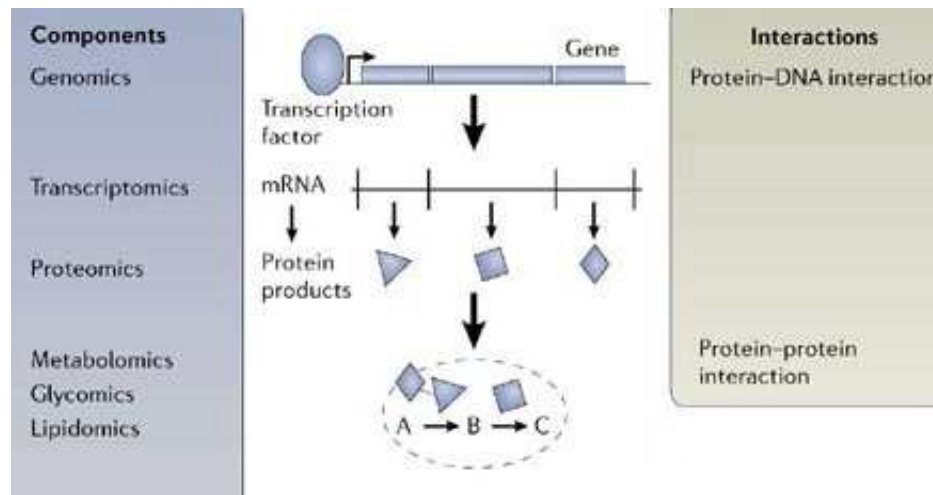


Figure 1.2: Genomic data is providing large-scale descriptions of nearly all components and interactions within the cell (adapted from [75]).

at different levels of the regulatory mechanism, three processes are explicitly listed below, followed by related types of data that can provide relevant information about these processes.

- **Transcription factor binding** Proteins that are transcription factors bind to genes/segments of DNA, cause changes in their expression and facilitate transcription.
- **Gene expression** Genes are transcribed into mRNAs, and the resulting mRNA abundance can indicate the active genes and their expression levels.
- **Protein-protein interaction** mRNAs are translated into proteins which perform cell functions. Some proteins that are transcription factors can again initialise gene expression.

Figure 1.2 depicts these three processes tracing the genetic information flow. As indicated in the adjacent boxes, the related data are classified into two categories, inter-

action data or component data. The interaction data specify links between molecular components while components data deal with the molecular content of the cell.

From the top, transcriptions factors (proteins) regulate and initiate transcription of mRNA from DNA. The processes that are responsible for generating and modifying these cellular components are generally dictated by molecular interactions, in this case, by protein-DNA interactions. These interactions can be described with the transcription factor binding data, which directly capture protein-DNA interactions in the first place. Then during transcription, genes are expressed and result in mRNAs. The presence and the relative abundance of resulting mRNA transcripts can be measured by the component data of the microarray gene expression data. After these mRNA are translated into proteins, protein-protein interactions are involved in translational processes as well as enzymatic reactions. Protein-protein interaction data can indicate how the end products interact and dictate cellular functions. This figure shows how genome-scale data conveniently provide rich information about the key processes occurring within the genome and proteome. Next, we review these representative data and the techniques that are used to generate them.

### **Transcription factor binding data**

(Transcription factor) Binding data directly identify interactions between proteins and DNA *in vivo*, particularly between transcription factors and their target genes. Such interactions fundamentally define the underlying regulatory network and reflect the binding kinetics of the constituent molecular species (genes and proteins). Binding data can be obtained with high-throughput technologies such as ChIP-Chip [98]. Still, binding is only a necessary condition for regulation. Many true positives at the binding level can be expected to be false positives at the regulatory level.

### **Microarray gene expression time data**

Microarray gene expression data record the levels of genes being expressed in order to determine the set of genes that are differentially expressed between two experimental treatments or conditions [114]. If microarray is used in expression profile experiments that are conducted at subsequent time points, the resulting data consist of gene expression level measurements taken at either uniformly or unevenly distributed time points. Gene expression data receive special attention in genomics research, both because of its rich information and genome-wide coverage. However, the data is prone to high degree of variability and noise due to inherent problems of the technique.

### **Protein-protein interaction data**

Protein-protein interactions play critical roles in dictating most cellular process, such as enzyme-complex formation and catalysis. In essence, information from protein-protein interactions not only potentially reveals sets of proteins that are involved in the same pathway, but also can be related to transcriptional regulation level in the sense that interacting proteins are often co-expressed and co-localised to the same sub-cellular compartment. Protein-protein interaction data can be obtained by the high-throughput scaling of technologies that exhaustively probes all the potential interactions within entire genomes, such as the yeast two-hybrid system [68]. However, these methods can suffer from high false positive and false negative rates owing to their inherent limitations [140].

The availability of these data makes it possible to understand gene functions and interactions. In this way, key gene or gene combinations can be found to explain spe-

cific cellular phenotypes which is the physical manifestation/change brought about by altered gene expression, e.g. disease susceptibility. However, it is important to recognise that high-throughput methods generally sacrifice specificity for scale, yielding many false positives and high-level noise in the data. Since gene expression time series data provide dynamic information about cell activities, they are the main focus of this thesis. The other two data types will be used in Chapter 4.

### 1.1.2 Microarray Data Acquisition

Of the three types of data, only microarray gene expression data analysed in this thesis are time series data. Microarray data are obtained from time series experiments to assess gene expression profiles in order to extract genomic information across time or under different experimental treatments. Gene expressions over time can be captured and recorded into a succession of numbers, on the scale of tens of thousands of genes. The dynamic information in time series data is useful in studying casual relations between time series, which are essentially equivalent to regulatory relationships between genes. Ideally, this will ultimately lead to mapping out the regulatory circuits in the genome [146].

Microarray is a high-throughput technology that can provide gene expression measurements for thousands of genes simultaneously. A microarray has a collection of DNA products printed onto a glass slide and each product is specific for an individual gene. mRNA from two biological samples are fluorescently tagged and hybridised simultaneously to probes on the array. Through competitive binding of these probes to the gene-specific DNA products a relative abundance of that gene within the two samples can be determined by capturing fluorescent signal information for each spot for

the two separately tagged probes. The underlying hypotheses are that the mRNA abundances in probes reflect the expression levels of the corresponding genes, and that the mRNA abundances decides, as a result of the detection, the strength of the signal under the excitation of a laser scanner, because abundant sequences will generate strong signals and rare sequences will generate weak signals. In other words, microarray takes snapshots of gene expression levels of all the genes in an organism.

To obtain microarray gene expression time series, microarray experiments are performed at different time points with either uniform or uneven intervals. Quantitative data are extracted from the resulting microarray images, normalised and processed into a gene expression matrix. Each row in this matrix describes the expression levels for one gene across time. Consequently, gene expression time series data are obtained as sequences of gene expression measured at successive time points at either uniform or uneven time intervals [114].

## 1.2 Statistical Inference for Functional Genomics

Statistical inference for functional genomics aims at integrating statistical inference methods and the understanding of functional mechanisms of the genome [40, 71]. To achieve the central goal of functional genomics, which is essentially extracting biologically relevant network topologies, various types of techniques such as clustering and network modelling can be utilised, so that problems can be systematically tackled. This section provides a summary of relevant literature and problem formularisations for the issues presented in this thesis.

### 1.2.1 Tight Clustering of Gene Expression Profiles

To deal with the large-scale gene expression data, clustering is usually the initial step towards biological inference for gene functions. Clustering aims to assign genes that share similar expression patterns into the same cluster. It provides an efficient way to extract information from large-scale gene expression data sets. Relevant genes can be screened out for the biological process under study, or possible functional relationships can be found among tens of thousands of genes on a microarray. The underlying assumption in clustering gene expression data is that co-expression indicates co-regulation, thus clustering should identify genes that share similar functions. This biological rationale is readily supported by both empirical observations and systematic analysis [17].

Given this promising direction, various clustering methods have been proposed to process the tremendous amount of microarray data, see [71, 72] for excellent reviews of current techniques. Looking at the prevalence of the many existing algorithms there may be no need to implement new ones. However, continuous development of the microarray technique brings new challenges on a regular basis. Moreover, many existing methods were adapted or even directly applied to gene expression data from conventional clustering algorithms [72], which may fail to meet current needs.

In particular, tight clustering arose recently from a desire to obtain tighter and potentially more informative clusters in gene expression studies [138]. Scattered genes with relatively loose correlations should be excluded from the gene clusters. Although various model-based clustering methods have been proposed, few of them address the need of obtaining tight and hence more biologically meaningful clusters. Objective methods that are specifically designed to address pertinent problems and new methods

are therefore essential. In Chapter 2, a new tight clustering method will be proposed to meet these requirements.

### 1.2.2 Clustering Validation Using Functional Annotation

With many clustering algorithms available, it is non-trivial to select one that can best tackle the challenges posed by the genomic data. Systematic formulation is therefore needed to prove the feasibility of clustering methods in this field. While it is still open to debate how a validation system should be constructed for gene expression clustering to verify the usefulness of the schemes, one promising direction is assessing the performance of an algorithm with existing biological knowledge. Outcomes from biological research have been gathered and translated into databases over decades, which provide specialised information to describe the functional profiles of genes. Exploiting information from these databases can facilitate integrative analysis of experimental results and existing knowledge, and further provide evidence for validation studies.

One of these databases, the Gene Ontology consortium (GO) [132], offers a wealth of complementary biological knowledge and is one of the most important ontologies for gene functions. Its structured vocabulary not only provides straightforward information about the gene functions, but it is also computationally accessible to quantify the relationships between genes. In essence, mappings are set up between genes and structured functional categories, and thereby annotating genes with a defined set of functions. Potentially, genes can be grouped according to their functional mappings to corresponding GO terms, which provides a good validation platform for a clustering method.

Consequently, many GO-driven methods have been proposed to establish func-



tional relationships between genes and, ultimately, to assess the quality of gene clusters [2, 29, 88]. However, none of them has systematically taken the structure information in GO into account. In Chapter 3, a GO-driven clustering validation index is proposed to make full use of the information provided by GO.

It is worth mentioning that another way of utilising GO knowledge in clustering analysis is to incorporate GO information into the clustering process in the hope of building more biologically meaningful clusters [22, 59, 129]. For example, it is proposed in [129] that the number of clusters can be determined by extracting Web-based knowledge to be used as input to their semi K-means algorithm. However, the effectiveness of this strategy greatly depends on the accuracy of knowledge about the organism under study. It neglects the fact that existing knowledge and the true patterns do not necessarily coincide and may as a result fail to discover the true biological patterns. One of the reasons for this contradiction originates from the incompleteness and false positives of biological databases. After all, an important goal for clustering is to identify novel functional annotations. Indeed, it is the desire of understanding the gap between statistical findings and current biological understanding that drives researchers.

### 1.2.3 Transcriptional Regulatory Network Reconstruction

A gene regulatory network is a collection of genes and gene interactions with each other indirectly through their RNA and protein expression products and with other substances in the cell, thereby regulating the rates at which genes in the network are transcribed. Reconstructing, or reverse-engineering, gene transcriptional regulatory networks can be defined as the process of identifying regulating interactions among

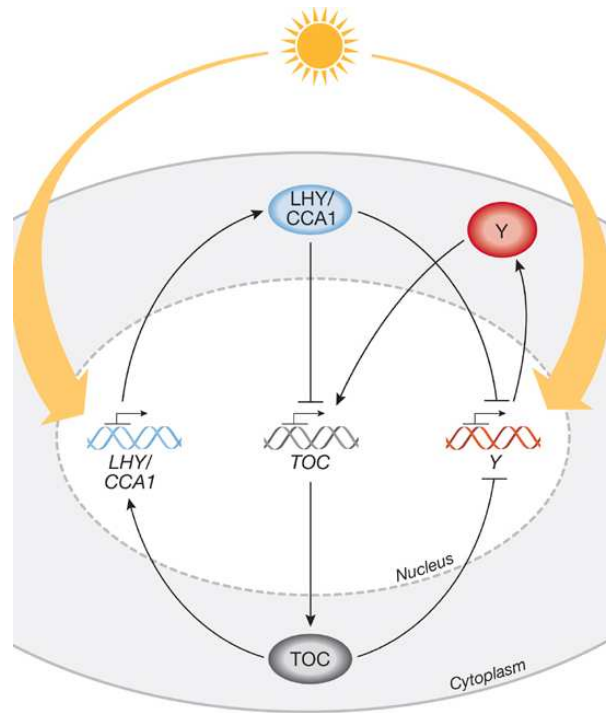


Figure 1.3: An Arabidopsis circadian gene network of six genes (adapted from Locke *et al* [87]). Circles are proteins and DNA segments represent genes.

genes from biological data. Nowadays, genomic, transcriptomic and proteomic data are in massive production. By using these high-throughput data, transcriptional regulatory activities are modelled on a genome-wide scale. Most importantly, gene network reconstruction helps clarify the role a gene plays in the transcriptional regulatory system, so that relevant genes, such as important transcription factors, can be screened out and chosen for further experimental manipulation to help consolidate the knowledge of the system under investigation.

As an example of a transcriptional regulatory network, a circadian gene network in *Arabidopsis Thaliana* is illustrated in Figure 1.3 as studied in [87]. One of the regulatory relationships in this network is the protein TOC, together with light derived input signals, activates the gene LHY/CCA1 transcription, while the protein LHY/CCA1 in

turn has regulatory effect on the transcription of the gene TOC.

Although gene expression data from microarrays are typically used for the purpose of transcriptional regulatory network reconstruction [3], it often lacks the desirable specificity and accuracy [11], since the information in the data is often entangled in a complex mixture of various types of noise. When more than one biological data source is available, integrative analysis is likely to offer significant advantages, and is currently a subject of ongoing research. As shown in Section 1.1, DNA, RNA, and protein interact with each other. The information from all three realms must be combined to bring full understanding of the global cellular structure. Integrative inference algorithms serve this purpose by exploiting, in addition to expression profiles, protein-protein interaction data, sequence data, protein modification data, metabolic data and more, in the inference process [3].

For integrative approaches, existing techniques have evolved from the simplest voting model [142] to more sophisticated Naïve Bayesian Networks [82, 120], and progressively to substantially more complex systems nowadays [86, 127]. However, so far there is no robust method that can be routinely applied to noisy and heterogeneous data and yet be efficient enough for handling gene expression time series [135, 161].

One of the issues in designing such an integrative system arises from the diverse formats of genomic data, which, in particular, is made explicit by the high-dimensional microarray time series. The study of inferring gene networks from microarray time series data alone fits well into classical theories of dynamic systems. However, for existing time series inference methods the microarray time series are very often too short to provide enough information about the regulatory relationships underneath concomitant behaviour changes.

Moreover, interactions between genes in a regulatory network do not necessarily

imply direct physical interactions, but can also refer to indirect regulations via proteins, metabolites and mRNA that have not been measured directly. Computational analysis that can differentiate between these two types of interactions are important to bring a better understanding to the underlying network structure. New, objective methods are needed in order to address these problems outside the boundary of classical theories. In Chapter 4, a new gene regulatory network inference method is proposed to make integrative inference from multiple data sources to increase predictive accuracy and to address these issues.

## 1.3 Thesis Overview

### 1.3.1 Thesis Contributions

Although biologists possess a basic understanding of the mechanisms regulating the flux and flow of information through this complex multidimensional regulatory system, they have not yet determined individual roles of most genes in the transcriptional system. This thesis presents computational tools to help infer gene functions, utilise current biological knowledge, and discover gene regulatory relationships. The approaches combine ideas from signal processing, graph theory, Bayesian models. In particular, it presents algorithms for gene clustering and network modelling, and solutions for inferring from large-scale, noisy and diverse genomic data. In this respect, it points out that efficiency, robustness and flexibility are the key to successful applications of statistical inference algorithms to this particular field of research.

One contribution of the thesis involves the discussion of advantages and limits in current gene expression clustering research. In particular, we address the emerging

problem in analysing gene expression data as discussed in Section 1.2.1, that gene clusters often need to be tight/small enough to provide strong evidence for gene function discovery. Although various clustering methods have been proposed, few of them address this need of obtaining tight clusters. At the same time, scattered genes with relatively loose correlations within clusters should be excluded from gene clusters. We point out that there is little work dedicated to this particular area of research in the literature. In response, a new tight clustering algorithm is proposed specifically aiming at the usually short gene expression time series.

The second contribution concerns utilising current biological knowledge for quantitative clustering validation. In Chapter 3, we analyse current progress in this field and bring up limits and challenges, before laying out a validation framework specifically designed for GO. Two validation indices have been developed, based on a new term-term distance defined within the realm of graph theory. Designed to overcome the challenges aforementioned in Section 1.2.2, the proposed validity indices take into account the variations in biological specificities for GO terms, the strength of relationships between terms, and the graphical structure of GO.

Another contribution involves proposing a new computational method for integrative analysis of heterogeneous data sources. Chapter 4 presents a Bayesian integrative framework for transcriptional regulatory network reconstruction with a Markov Random Fields [80] component that applies the tight clustering method proposed in Chapter 2. A stochastic process for parameter estimation is designed to achieve robustness to noise, yet the system remains efficient enough to facilitate large-scale analysis. This chapter not only addresses the issue of integrating different formats of genomic data by providing a simple yet effective solution, but also reveals diverse characteristics of different types of biological data.

### 1.3.2 Thesis Organisation

Chapter 2, 3, and 4 constitute the three core analytical chapters in this thesis. Each chapter has its individual section of literature review, emphasising on the research gaps in the current literature. In an attempt to bridge these gaps, a solution is proposed and demonstrated to be effective in the experimental section, independently.

Nevertheless, all chapters are connected in one way or another. The partial mixture model-based clustering algorithm proposed in Chapter 2 serves as a preliminary step towards inference and is used throughout the thesis when necessary. Complementary to Chapter 2, Chapter 3 introduces a Gene Ontology-driven validation method, providing evidence of the superior performance of the partial mixture clustering algorithm. Moreover, it provides useful insights into the complex structure of Gene Ontology. Chapter 4 constitutes the gene network inference part of research. In Chapter 4, an integration framework for combining different biological sources is proposed for transcriptional regulatory gene network reconstruction. Such a network is useful in discovering relevant network structure and identifying important genes in certain cellular process.

Chapter 5, the concluding chapter, summarises the finding of the studies in this thesis, while providing insights into their implications and impacts to this field. It reviews the goals set in Chapter 1, objectives raised and solutions presented in the subsequent chapters, and points out promising directions for further research.

## **Chapter 2**

# **Partial Mixture Model for Tight Clustering Gene Expression**

### **2.1 Introduction**

With the advances of high-throughput microarray techniques, gene expression data clustering has been an active research area. Gene expression clustering aims to reveal groups of genes that share similar functions in the biological pathways. In particular, consider gene expression time series experiments, where the data are made up of tens of thousands of genes, each with measurements taken at either uniformly or unevenly distributed time points. For such large-scale data sets as the gene expression time series, clustering provides a good initial investigation tool, which ultimately leads to biological inference.

In this chapter, we review previous advances, discuss existing problems and propose a novel clustering algorithm specifically targeting gene expression time series. Some of the materials in this chapter have appeared before in [157]. The rest of this

chapter is outlined as follows.

In Section 2.2, we first review probabilistic models on which some popular clustering methods are based, and the parameter estimation methods that are routinely applied, in order to understand the intrinsic problems in existing clustering methods. Then through a discussion of current research trends, we show that conventional clustering algorithms cannot be simply adapted and applied to this field. In contrast, innovative and objective set ups are needed to tackle new problems in high-throughput genomic data in the hope of revealing biologically meaningful results.

In response to the existing problems and new challenges, a partial mixture model teamed with a minimum distance estimator is formulated for gene expression tight clustering in Section 2.3. The inherent robustness of the minimum distance estimator is experimentally proved, which makes it a powerful tool for outlier detection in model-based clustering. In the comparative experiments in Section 2.4, both biological and statistical validations for the proposed method are conducted on a simulated data set and two real gene expression data sets. The superior performance of the proposed method is confirmed by both biological and statistical validity indices. Moreover, the experimental results show that the tight clusters obtained by our proposed method are more biologically informative. This further proves the suitability of the proposed method in this field.

The study concludes by providing new biological hypothesis from the integrative analysis of the machine learning results and current biological knowledge. We show that tight clustering is capable of generating more profound understanding of the data set, well in accordance to established biological knowledge. It also provides new interesting hypotheses from the interpretation of clustering results. In particular, we provide biological evidence that scattered genes can be relevant and are interesting subjects for



study, in contrast to prevailing opinion.

## **2.2 Existing Methods and Future Needs**

Various model-based methods for clustering gene expression data have been proposed following the advances of microarray technique. Among them, finite mixture model methods are the most popular [63, 152]. Finite mixtures of distributions have offered a sound mathematical-based approach to statistical modelling [96].

A typical routine using these methods consists of two stages. First, a finite mixture model of the form  $p(x) = \sum_{i=1}^K w_i p_i(x, \theta)$  for a random variable  $x$  is designed.  $w_i$  is the proportion of the corresponding density  $p_i(x, \theta)$  with parameters  $\theta$ . Assuming that there is an underlying true model/density, three sets of parameters need to be estimated or explicitly specified: the number of clusters  $K$ , the proportions of clusters  $w$  and the parameter settings  $\theta$  for the densities. Then, the optimal parameters for the model are systemically found, so that the fitted model/density is as close to the true model/density as possible.

For modelling time series,  $p_i(x, \theta)$  is usually designed with a linear model to capture the dynamics in time series. Two of the most popular linear models are described here.

### **2.2.1 Linear Models**

In order to design an appropriate model, continuous representation of gene expression time series are preferred to capture the system dynamics. Many existing models for fitting gene expression time series fall into one or more of these categories: the spline regression models [4, 16, 63], the mixed effects models [92, 101] and the autoregressive models [151]. Next, we briefly review some of these representative models in the

literature.

### 2.2.1.1 Spline model

Spline models have received special attention in the clustering community for their desirable properties. For example, the use of piecewise low-degree polynomials results in smooth curves and avoids the problems of overfitting. Take the model in [4] as an example, cubic polynomials with B-spline basis are used for fitting gene expression time series data. Cubic polynomials are the lowest degree polynomials that allow for a point of inflection. The advantage of B-spline lies in that the degree of the polynomials is independent from the number of points and that curve shape is controlled locally.

A cubic spline consisting of  $\iota$  parameterised polynomials can be formulated as

$$y(t) = \sum_{i=1}^{\iota} A_i S_i(t), \quad (2.1)$$

where  $y$  is a vector of data,  $A_i$  are the coefficients and  $S_i$  are the polynomials.  $t$  is the parameter which, in the case of time series analysis, refers to time.

For the application to gene expression time series data, it is desirable to use B-spline basis to obtain smoothing spline, as smoothing spline use fewer basis coefficients than observed data points thus avoiding overfitting. Suppose observations are made at  $m$  time points, this imposes the constraint  $\iota < m$ . Using the Cox-deBoor recursion formula [110], the B-spline basis can be calculated as

$$b_{j,0}(t) = \begin{cases} 1, & \text{if } s_j \leq t < s_{j+1} \\ 0, & \text{otherwise,} \end{cases} \quad (2.2)$$

$$b_{j,k}(t) = \frac{t - s_j}{s_{j+k} - s_j} b_{j,k-1}(t) + \frac{s_{j+k+1} - t}{s_{j+k+1} - s_{j+1}} b_{j+1,k-1}(t). \quad (2.3)$$

As the order of the basis polynomials,  $k$  is 4 for cubic polynomial.  $s_j$  are the knots where  $j$  is in the range of  $[1, \iota + k]$ . As splines are piecewise polynomials, the abscissa values of the join points where the polynomials join are called knots. Knots give the curve freedom to fit more closely to the data. The use of knot vector  $s$  particularly suits microarray data analysis, since it can be defined to be either uniform or unevenly spaced. According to the purpose of the microarray experiments, it is sometimes desirable to place more knots where biological activity is intense, instead of using the uniform knot vector as in [4].

For applications, take the mixture model in [4] as an example. Let  $Y$  denotes gene expression data of  $n$  genes  $\{y_i | i = 1 \dots n\}$ , the mixture model is also a mixed effect model with both cluster specific and gene specific coefficients

$$y_i = S_i(\mu_j + \gamma_{i,j}) + \varepsilon_i. \quad (2.4)$$

$\mu_j$  denotes the average value of the spline coefficients for genes in class  $j$ , and  $\gamma_{i,j}$  denotes the gene specific variation coefficients, depending on the class assignment  $j$ .  $\varepsilon_i$  is Gaussian noise. Both  $\gamma_{i,j}$  and  $\varepsilon_i$  are normally distributed with mean zero and variances  $\Gamma_j$  and  $\sigma^2$ , respectively.

### 2.2.1.2 Autoregressive model

Suppose  $Y = \{y_i | i = 1, 2, \dots, n\}$  is a multivariate stationary time series of  $n$  variables and  $t$  time points. A  $p$ -order vector autoregressive model specifies the value of a variable at a time point  $t$  as formulated in Eq.(2.5).  $Y(t)$  is a linear combination of a constant/mean

value, the past of the multivariate time series, and noise

$$Y(t) = B + A \sum_{u=1}^p Y(t-u) + \varepsilon(t). \quad (2.5)$$

$B$  is a constant matrix of size  $n \times t$ .  $\varepsilon$  consists of vectors of residuals  $\{\varepsilon_i | i = 1 \dots n\}$ , each is assumed to be zero mean noise with variance  $\sigma_i^2$ .  $A$  is a  $n \times n$  coefficient matrix representing the dynamic structure. A special case of the  $p$ -order autoregressive process, the first-order autoregressive model is often considered when analysing microarray data for the sake of simplicity [81, 102, 151]

$$Y(t) = B + AY(t-1) + \varepsilon(t). \quad (2.6)$$

When  $A$  is a constant matrix, this model assumes homogeneity across time. The parameters are often estimated by optimisation methods such as the maximum likelihood estimator (MLE) [63, 108, 141].

### 2.2.2 Parameter Estimation

For the task of parameter estimation, the maximum likelihood estimator (MLE) is one of the most extensively used statistical estimation techniques in the literature. For a variety of models, maximum likelihood functions [63, 101, 141] have been applied for estimating parameters of probability distributions. The solution often involves maximising the likelihood over each parameter by iteratively applying the expectation-maximisation (EM) algorithm [35]. Examples abound [4, 63, 92, 94, 101].

The EM algorithm alternates between inference about the hidden variables (the expectation step) and maximal likelihood estimation of the model parameters (the max-

imisation step). The expectation step in EM first uses temporary data to represent a reasonable guess for the hidden variables. Then the parameter estimation proceeds as if the data is complete, maximising a likelihood function for the parameters. Once a solution for the parameter estimates is produced, it is used to place the temporary data values with better guesses. The two-step process is then repeated again until convergence, i.e., when the difference between the parameters updates is smaller than a predefined value.

For example, EM is used to determine the maximum likelihood estimation for the model in [4] (see Eq.(2.4)). Cluster memberships are treated as missing data. The optimisation problem can be decomposed in the following way, assuming  $\gamma_i$  has been observed:

$$\begin{aligned}
 & p(Y, \gamma_{i,j} | \Gamma, \sigma^2, \mu) \\
 &= p(Y | \gamma_{i,j}, \Gamma, \sigma^2, \mu) p(\gamma_{i,j} | \Gamma, \sigma^2, \mu) \\
 &= \prod_i \prod_j Z(j|i) \frac{1}{(2\pi)^{n_i} \sigma^{n_i}} \exp\left[-\frac{1}{2\sigma^2} (Y_i - S_i(\mu_j + \gamma_{i,j}))^T (Y_i - S_i(\mu_j + \gamma_{i,j}))\right] \times \\
 & \quad \frac{1}{(2\pi)^q \Gamma_j^{1/2}} \exp\left[-\frac{1}{2\Gamma_j} \gamma_{i,j}^T \gamma_{i,j}\right],
 \end{aligned} \tag{2.7}$$

where  $Z(j|i)$  is a binary indicator variable that assigns each gene to exactly one class.

The E-step finds the probability of each gene  $i$  belonging to cluster  $j$ ,  $p(j|i)$ ,

$$p(j|i) = \frac{p_j p(Y_i | \gamma_{i,j}, \Gamma_j, \sigma^2, \mu_j)}{\sum_k p_k p(Y_i | \gamma_{i,k}, \Gamma_k, \sigma^2, \mu_k)}. \tag{2.8}$$

The M-step maximises the parameters with respect to the probability  $p(j|i)$ . And

at the end the cluster probability  $p_j$  are updated through

$$p_j = \frac{1}{n} \sum_i^n p(j|i). \quad (2.9)$$

The two-step process is then repeated until convergence is reached. Each gene  $i$  is then assigned to class  $j$  that maximises  $p(j|i)$ .

### 2.2.3 Limitations of Existing Methods

It is observed that model-based approaches generally achieve superior performance to many others [46, 63, 133, 153]. Nevertheless, current methods generally rely on correct model assumption. For example, the autoregressive model as described in Section 2.2.1.2 requires Markov property and stationarity [94]. The former requirement may not hold for some time series data. Stationarity means the system that generates time series should be time invariant. Thus the temporal structure of the data and the length of sampling intervals are not considered in this approach.

Also, rigorous statistical inference is needed for the estimation of model parameters. The parametric nature of existing methods requires an optimisation process which might be time consuming. The initial values to start the optimisation often need fine tuned. To be specific, the problem with the quasi-Newton type of optimisation methods [31] is that the quantities can be estimated only when they satisfy some constraints, while with EM, some parameters have to be explicitly specified and others have to be initialised. For example, in [4, 101, 151] the number of clusters  $K$  has to be known *a priori*, which is not practical in microarray data analysis. Moreover, the existence of local optima of the likelihood function and the requirement for an initial configuration mean that several runs are needed before a satisfactory clustering outcome is produced.

Apart from the aforementioned issues, clustering algorithms may have other inherent problems. For example, SplineCluster [63] is an efficient hierarchical clustering program based on nonlinear regression splines. The use of nonlinear spline basis can accommodate non-stationary time-dependence and unequal intervals in the data. Starting from singleton clusters, the idea is to successively merge clusters based on a potential function to form a dendrogram. The algorithm is efficient and straightforward to visualise. However, as a common problem to all hierarchical clustering methods, the broadest clusters often contain many scattered genes and can sometimes be hard to interpret, as later merges often depend on aggregated measures of clusters.

In summary, existing methods have their inherent problems. Multivariate Gaussian models [44] ignore the time order of gene expression and therefore cannot account for the correlation structure in time series data [94]. Spline models [4, 63, 92, 94] and autoregressive models [151], such as the ones presented in Section 2.2.1.1 and Section 2.2.1.2, generally apply EM for parameter estimation, and are computationally expensive for large data sets. Moreover, there has been extensive use of maximum likelihood estimator (MLE) [78] for model parameter estimation. By contrast, the minimum distance estimator (MDE) [10] has been largely ignored.

### 2.2.4 Emergence of Tight Clustering

Intuitively, tight clustering refers to methods that can be built upon an existing partition to obtain core patterns that are more easily interpretable. The initial partition can be obtained either empirically or by using generic algorithms such as the K-means algorithm. Only clusters of closely related genes are then separated from these clusters, leaving scattered genes out. As a result, more information can possibly be revealed

from tight clusters. For example, being in the same tight cluster is strong evidence that the genes share similar functions. Or, if genes in one functional category are allocated into different tight clusters, one may pursue possible explanation by looking into these clusters. One possible result of such investigation is that some genes have unknown functions that affect their expression patterns, hence leads to new gene function discovery.

In this sense, to obtain tight clusters, some genes should be classified as scattered genes, if forcing them into clusters will only disturb biologically relevant patterns. Indeed, the issue of scattered genes has received more attention only recently [71, 138]. Currently, there are few methods to deal with scattered genes with respect to the analysis of gene expression time series.

To the best of our knowledge, the work of [138] is the first to address the problem about tight clustering. But it relies on heavy computation due to the nature of random resampling. For the methods that address the problem of scattered genes, one popular method is MCLUST [44, 153]. MCLUST is an unsupervised method based on multivariate Gaussian models. The models are characterised by their geometric features for the clusters: shape, orientation and volume. Each time the best models are selected for the data set being clustered and then the model parameters are estimated by EM. It was proposed in [45] that outliers can be modelled by adding a Poisson process component in the mixture model. A recent implementation of MCLUST [47] allows an additional component of homogeneous Poisson process for modelling scattered genes/noise. This method relies on correct model specification and the robustness of the parameter estimator.



## 2.3 Proposed Tight Clustering Method

When analysing gene expression time series data, special attention needs to be paid to the following issues:

- **Number of clusters:** The main difficulty about the model-based methods concerns the number of clusters  $K$ , which has to be specified most of the time. It is particularly problematic for microarray data, which may be evenly distributed in the gene expression space and thus may not have any straightforward solution featuring isolated clusters.
- **Scattered genes:** Recently, it has been proposed to allow a noisy set of genes not being clustered [138]. In microarray experiments, it is generally expected that many genes could show uncorrelated variations and are unrelated to the biological process under investigation. Forcing these genes into clusters will only introduce more false positives, resulting in distorted clusters and difficulty in interpretation. It is later experimentally verified that methods that allow for scattered genes give better accuracy and robustness [133].
- **Tight clusters:** It is suggested that tight clusters are often more biologically informative, typically of size 20-60 genes [138]. Conventional methods produce large and loose clusters, while biologists often need to conduct research on smaller groups of closely related genes. Therefore, tight clustering has been proposed for obtaining smaller and tighter clusters from gene expression data.

In this section, we present our partial mixture model algorithm to address the above challenges in a semi-parametric fashion. Built upon the advantages of MDE and partial modelling, the algorithm performs tight clustering which naturally incorporates replication information and allows a set of scattered genes to be left out.

To relieve the system of the tedious parameter optimisation process, our proposed partial mixture model is based on the MDE instead of the MLE. There are many unique features of MLE, including its efficiency. However the practical deficiencies of MLE, besides those issues with its optimisation, are the lack of robustness against noise and its sensitivity to the correctness of model specification. We discuss in this chapter the performance of the appealing alternative, MDE, which is less explored in this field. Inspired by the work of [117, 118], MDE is used to relax the system's dependence on parameter optimisation. MDE provides robust estimation against noise and outliers, which is especially appropriate for gene expression data analysis, where data are often noisy. To show the improvement in performance offered by the new method, we compare the proposed method to SplineCluster and MCLUST in the experiments.

### 2.3.1 Minimum Distance Estimator (MDE)

Given a density function  $f(\cdot)$ , its corresponding parameters  $\theta$  and  $n$  variables of interest  $x_i, i = 1, 2, \dots, n$ , we aim to find the optimal parameters  $\hat{\theta}$  to approximate the true parameters  $\theta_0$  by minimising the integrated squared difference

$$d(f(\theta), f(\theta_0)) = \int [f(x|\theta) - f(x|\theta_0)]^2 dx, \quad (2.10)$$

which gives

$$d(f(\theta), f(\theta_0)) = \int f(x|\theta)^2 dx - 2 \int f(x|\theta)f(x|\theta_0)dx + \int f(x|\theta_0)^2 dx. \quad (2.11)$$

The last integral  $\int f(x|\theta_0)^2 dx$  is a constant with respect to  $\theta$ , thus can be ignored. The second integral can be obtained through kernel density estimation [105]. Therefore,

the MDE criterion simplifies to

$$\hat{\theta} = \arg \min_{\theta} \left[ \int f(x|\theta)^2 dx - \frac{2}{n} \sum_{i=1}^n f(x_i|\theta) \right]. \quad (2.12)$$

There are many interesting features of MDE. First of all, it comes with the same robustness as all other minimum distance techniques [52, 95, 104, 159]. Secondly, MDE approximates data by making the residuals as close to normal in distribution as possible [52, 95, 159], which will turn out to be very useful for the model set up to be described later. These features will be further explained and illustrated in the experiments. We will also illustrate derivation of the MDE criterion for parameter estimation for our partial regression algorithm. Owing to space restrictions, some discussions, results and elaborations have been relegated to Appendix A.

### 2.3.2 Weighted Mixture Model with MDE

In principle, the finite mixture model methodology assumes that the probability density function,  $f(x|\theta)$ , can be modelled as the sum of weighted component densities. The weights are often constrained to have a sum of 1. It is revealed later that this constraint is not necessary. More flexible models can be obtained by relieving the system from this constraint. A weighted Gaussian mixture model has the form:

$$f(x|\theta) = \sum_{k=1}^K w_k \phi(x|\mu_k, \sigma_k^2), \quad w_1 + w_2 + \dots w_K = 1, \quad (2.13)$$

where  $\phi$  is the Gaussian density function,  $\mu, \sigma$  are the mean and standard deviation,  $K$  is the number of components, and  $w_k, k = 1, 2, \dots, K$  are the weight parameters. However, by relieving the constraint of  $\sum_{k=1}^K w_k = 1$ , the system can be extended for overlapping

clustering inference [117] since the sum of the amount of data being modelled in all clusters can exceed the total amount of data. Later, we will further prove that the amount of modelled data can also be less than the total amount of data. In all cases,  $w_k$  indicates the proportion of data points that are allocated in the  $k$ th component. Let  $g_K(x|\theta)$  be the part in Eq.(2.12) to be minimised for a  $K$ -component mixture model, we have

$$g_K(x|\theta) = \int f(x|\theta)^2 dx - \frac{2}{n} \sum_{i=1}^n f(x_i|\theta). \quad (2.14)$$

On the other hand,

$$\begin{aligned} \int \phi(x|\mu, \sigma^2)^2 dx &= \int \left[ \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \right]^2 dx \\ &= \frac{1}{2\sigma\sqrt{\pi}} \int \frac{1}{\sqrt{2\pi}\frac{\sigma}{\sqrt{2}}} \exp\left(-\frac{(x-\mu)^2}{2(\frac{\sigma}{\sqrt{2}})^2}\right) dx \\ &= \frac{1}{2\sigma\sqrt{\pi}}. \end{aligned} \quad (2.15)$$

And from [143, Section 2.6],

$$\begin{aligned} &\int \phi_1(x|\mu_1, \sigma_1^2) \phi_2(x|\mu_2, \sigma_2^2) dx \\ &= \phi(\mu_1 - \mu_2 | 0, \sigma_1^2 + \sigma_2^2) \int \phi(x | \frac{\sigma_1^2 \mu_2 + \sigma_2^2 \mu_1}{\sigma_1^2 + \sigma_2^2}, \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2}) dx \\ &= \phi(\mu_1 - \mu_2 | 0, \sigma_1^2 + \sigma_2^2). \end{aligned} \quad (2.16)$$

By combining Eq.(2.13), (2.15) and (2.16), we have

$$\begin{aligned}
& \int f(x|\theta)^2 dx \\
&= \int \left( \sum_{k=1}^K w_k^2 \phi(x|\mu_k, \sigma_k^2) + \sum_{k=1}^K \sum_{l=1}^K w_k w_l \phi(x|\mu_k, \sigma_k^2) \phi(x|\mu_l, \sigma_l^2) \right) dx \\
&= \sum_{k=1}^K \frac{w_k^2}{2\sqrt{\pi}\sigma_k} + \int \sum_{k=1}^K \sum_{l=1}^K w_k w_l \phi(x|\mu_k, \sigma_k^2) \phi(x|\mu_l, \sigma_l^2) dx \\
&= \sum_{k=1}^K \frac{w_k^2}{2\sqrt{\pi}\sigma_k} + \sum_{k=1}^K \sum_{l=1}^K w_k w_l \phi(\mu_k - \mu_l|0, \sigma_k^2 + \sigma_l^2).
\end{aligned} \tag{2.17}$$

Thus from Eq.(2.14) and (2.17), the distance for the  $K$ -component Gaussian mixture model can be expressed as

$$\begin{aligned}
& g_K(x|\theta) \\
&= \sum_{k=1}^K \frac{w_k^2}{2\sqrt{\pi}\sigma_k} + \sum_{k=1}^K \sum_{l=1}^K w_k w_l \phi(\mu_k - \mu_l|0, \sigma_k^2 + \sigma_l^2) - \frac{2}{n} \sum_{i=1}^n \sum_{k=1}^K w_k \phi(x_i|\mu_k, \sigma_k^2).
\end{aligned} \tag{2.18}$$

$g_K(x|\theta)$  is a closed-form expression, whose minimisation can be performed by a standard nonlinear optimisation method. We use a Newton-type algorithm [36] implemented in R as a function `nlm()`.

For example, a one-component model has the following MDE criterion

$$\begin{aligned}
\hat{\theta} &= \arg \min_{\theta} [g_1(x|\theta)] \\
&= \arg \min_{\theta} \left[ \frac{w^2}{2\sqrt{\pi}\sigma} - \frac{2w}{n} \sum_{i=1}^n \phi(x_i|\mu, \sigma^2) \right].
\end{aligned} \tag{2.19}$$

We aim to further relieve the system from the constraints posed by the weight parameters, whilst keeping its weighted-component structure. In the next section the idea of

partial modelling is presented. It originated from the fact that incomplete densities are allowed [7], so the model will be fitted to the most relevant data.

### 2.3.3 Partial Mixture Model with MDE (PMDE)

The weight parameters are of particular importance in a partial mixture model. They allow the model to estimate the component/components, while their value indicates the proportions of fitted data, so the rest of the data can be treated as scattered genes/outliers. This approach is first described in [117] for outlier detection. It was suggested in [71] that by forcing a large scaling parameter in one of the components in the mixture, scattered genes can be accommodated in this component. However, partial modelling provides a more flexible alternative approach, as described later.

Although it is suggested in [117] that the unconstrained mixture model can be applied for clustering, through our experiments it is clear that if the data overlap to a certain degree, all components will converge to the biggest component as a result of model freedom. Moreover, it is impractical to formulate the criterion in the form of Eq.(2.18) when it comes to implementation. Instead, we solve the problem by taking advantage of the one-component model to formulate our clustering algorithm.

#### 2.3.3.1 The spline regression model

To provide continuous representations of gene expression time series profiles, a linear regression model with nonlinear cubic spline bases is set up. The linear regression model is capable of capturing the inherent time dependence, while the nonlinear spline bases help accommodate the underlying stochastic process in the data. The advantage of using cubic spline lies in the fact that degree of the polynomials is independent of

the number of points and that the curve shape is controlled locally.

Let  $Y$  be the gene expression data matrix of size  $n \times m$ , with  $n$  the number of genes to be modelled and  $m$  the number of time points.  $Y_i$  can be modelled as

$$Y = \alpha + \mathbf{Q}\beta + \varepsilon. \quad (2.20)$$

$\mathbf{Q}$  is the design matrix of size  $nm \times q$  consisting of a linear combination of cubic B-spline basis functions as described in Eq.(2.2) and (2.3), with  $q$  being the number of knots. The error term  $\varepsilon$  represents the residuals taken as a weighted Gaussian distribution  $w \cdot N(0, \sigma_\varepsilon^2)$ .  $\alpha$  is the intercept and the  $q$ -vector  $\beta$  are the regression coefficients.

As stated before, the useful feature of MDE is that it fits data in such a way that the residuals are close to normal, so that the residual can be modelled by a normal distribution. Therefore, our model becomes

$$\varepsilon = Y - \alpha - \mathbf{Q}\beta. \quad (2.21)$$

Given Eq.(2.13), (2.15) and (2.21), the one-component PMDE fit for this model has the form of

$$\begin{aligned} \hat{\theta} &= \arg \min_{\theta} \left[ \int (w\phi(\varepsilon|0, \sigma_\varepsilon))^2 d\varepsilon - \frac{2}{n} \sum_{i=1}^n w\phi(\varepsilon_i|0, \sigma_\varepsilon^2) \right] \\ &= \arg \min_{\theta} \left[ \frac{1}{2\sqrt{\pi}} w^2 \sigma_\varepsilon^{-1} - \frac{2w}{n} \sum_{i=1}^n \phi(\varepsilon_i|0, \sigma_\varepsilon^2) \right], \end{aligned} \quad (2.22)$$

where  $\theta = \{w, \alpha, \beta_1, \dots, \beta_q, \sigma_\varepsilon\}$ , and  $\phi$  is the density of a normal random variable. Altogether there are  $q + 3$  parameters to be estimated.

The number of knots  $q$  determines the degree of smoothing and can be chosen

according to the quality of data. Detailed discussion of this issue are skipped in this thesis. Interested readers are referred to [48]. In summary, this spline regression model captures the inherent time dependencies among data, where the error term is of particular importance as it can pick up the noise.

The proposed algorithm is designed specifically for gene expression clustering, taking into consideration the issues raised in Section 2.2.3. The knot vector in the spline model can accommodate uniform or unevenly spaced time points, as noted in Section 2.2.1.1. MDE, with its robustness, is excellent in detecting outliers/scattered genes. The number of clusters is determined by the algorithm itself, by setting a stopping criteria for it.

### 2.3.3.2 The stopping criteria

A statistical measure of partition quality, the Calinski and Harabasz (CH) index [21] as formulated in Eq.(2.23), is used to design a stopping criteria for the proposed algorithm. The CH index is given as

$$CH(K) = \frac{BSS(K)/(K-1)}{WSS(K)/(n-K)}, \quad (2.23)$$

where  $BSS(\cdot)$  and  $WSS(\cdot)$  are the between-cluster and within-cluster distances defined as

$$BSS(K) = \frac{1}{2} \sum_{l=1}^K \sum_{x_i \notin C_l, x_j \in C_l} d^2(x_i, x_j), \quad (2.24)$$

$$WSS(K) = \frac{1}{2} \sum_{l=1}^K \sum_{x_i, x_j \in C_l} d^2(x_i, x_j). \quad (2.25)$$



$x_i$  and  $x_j$  stand for the  $i$ th and the  $j$ th variables. Squared Euclidean distance is used for distance measurement  $d^2(x_i, x_j)$ .  $C_l$  in Eq.(2.24) and (2.25) stands for the  $l$ th cluster. The idea behind the CH measure is to compute the pairwise sum of squared errors (distances) between clusters and compare that to the internal sum of squared errors for each cluster. In effect, it is a measure of between-cluster dissimilarity over within-cluster dissimilarity. The optimum clustering outcome should be the one that maximises the CH index in Eq.(2.23).

### 2.3.4 Experimental Validation of MDE with partial modelling

The main feature of our model is its ability to identify the key component, if any, and a set of outliers, in order to find the data structure. Therefore, a feasible parameter estimator is of paramount importance. We empirically validate our points about the nature of partial modelling and MDE through fitting four simple simulated data sets. The performance of both MDE with partial modelling and MLE with a one-component spline regression model ( $K=1$ ) is compared in terms of data fitting accuracy and robustness. All data sets are generated by sine functions, modelling cyclic behavior of genes, which are widely employed in the literature [92, 152]. Time series length is 25 time points which is a typical number of microarray experiments. Gaussian noise is added to all data. The number of knots for both spline models is chosen to be 15 according to the stepwise selection criterion of knots in regression splines [123], to allow for flexibility in curves while avoiding overfitting. Surprisingly, superior performance was achieved for the PMDE fits even on such simple data sets.

We begin with simulating the situation when the number of components (3) in the data is seriously underestimated, as illustrated in Figure 1(a). Three components are

generated from three sine waves simulating gene expression data of three clusters. The components comprise 60%, 20% and 20% of the data, respectively. The MDE with partial modelling fit is highlighted by the pink line and the MLE fit is blue. MDE with partial modelling locates the major component, while MLE is biased to all data. MDE with partial modelling appears to be superior to MLE in such a scenario. The fact that the MDE with partial modelling can find the key component without compromising the others suggests a solution to the vexing problem when the number of components is unknown, which is often the situation in gene expression clustering. Histograms of residuals from both fits are plotted in Figure 2.1(b) and (c) to prove that MDE with partial modelling fit the data in such a way that the residuals are close to normal.

More data sets shown in Figure 2.1 (d)-(f) are used to compare the performances of MDE and MLE in different scenarios. When there are two components of entirely opposite behaviors, we can see from Figure 2.1(d) that the MLE fit is almost flat, while MDE fits the larger component (60% of the data). The situation where lots of outliers are present is simulated in Figure 2.1(e), where the major component has 60% of the data and the rest (40%) are generated from three different sine waves. MDE demonstrates its robustness by capturing the major component, while MLE is biased. However, in the case of two clusters of exactly equal size as shown in Figure 2.1(f), MDE fails, as it is designed to capture only one component but now cannot decide which one to fit. This can be solved by using a multi-component model.

From these examples, it is observed that MDE has the ability to identify the relevant fraction of data and distinguish it from outliers, while MLE blurs the distinction by accounting for all data. This is of great value for massive data sets, when the data structure is unclear and lots of outliers are present. The smoother fits of the proposed MDE than that of MLE manifest the fact that the former is more robust against noise.

### **2.3 Proposed Tight Clustering Method**

---

All these suggest MDE a promising tool for microarray data analysis. Interested readers are referred to Appendix A, for comparison of the two estimators on theoretical ground.

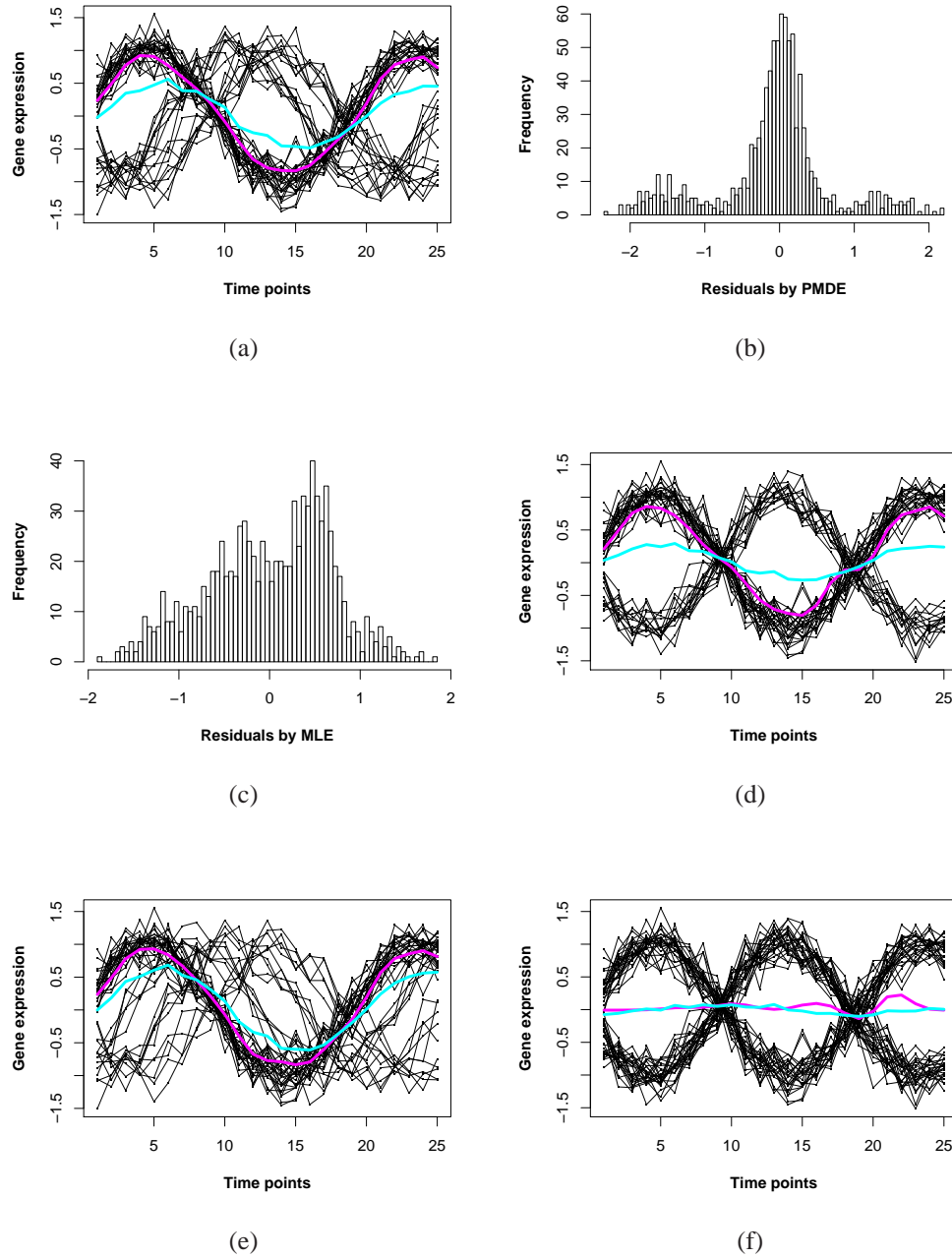


Figure 2.1: Comparing MDE and MLE by data fitting and their residual histograms. (a) MDE fit (pink line) and MLE fit (blue line) to simulated data generated from three sine waves; (b) Histogram of residuals by MDE; (c) Histogram of residuals by MLE; (d) MDE fit (pink line) and MLE (blue line) fit to simulated data generated from two sine waves; (e) MDE fit (pink line) and MLE (blue line) fit to data with many outliers; (f) MDE fit (pink line) and MLE (blue line) fit when two components are of same size.

### 2.3.5 The PMDE Clustering Algorithm

Tight clustering, by definition, builds compact clusters upon an existing partition. The initial partition, if not available, can be obtained by some empirical knowledge or heuristic clustering methods such as K-means. Given an initial partition, the clustering procedure is formulated as in Algorithm 1.

In the initialisation step of the algorithm, an existing partition of a data set is provided

---

**Algorithm 1** Partial Regression Clustering

---

**Require:** Initialisation: an initial partition is obtained.

**repeat**

1. Fit partial regression model to each of the clusters;
2. Identify potential outliers according to a tightness threshold  $\nu$  and discard them from the clusters;
3. For all outliers, fit partial regression model to form a new cluster;

**repeat**

4. For all genes re-evaluate distances to all existing spline regression models, assign them to the closest one;
5. Fit partial regression models to all clusters;
6. Calculate CH value based on current partitions;

**until** the clustering quality measured by CH value fails to improve;

7. Take the partition with highest CH value;

**until** no partial regression model can be fitted to the outliers;

8. Label all outliers as scattered genes.
- 

as input. The tightness threshold,  $\nu$ , controls the tightness and the number of refined clusters produced by the algorithm as output. It is defined as the reciprocal of the weighted mean variance of the clusters of the initial partition. Therefore, the greater the threshold is (i.e., the smaller the variance is), the tighter the clusters become and the more the clusters are formed. The weights are determined in proportion to the size of the clusters. In the main loop, after each new cluster is generated, all data points are reassigned in the gene redistribution loop, so the resultant clusters should be of reasonable size. The rationale supporting our design is based on the features of partial

modelling and robustness of the MDE estimator, which we believe is able to find the relevant components in the data, while not being distracted by outliers. The residuals, as a natural byproduct of model fitting, can be used as the distance between data points and spline regression models. The effectiveness of the algorithm depends on the model normality. Often gene expression data are transformed during pre-processing so that data normality holds approximately. When the model normality holds approximately, clusters can be found.

In this framework, we use deterministic class assignment during the clustering process. Stochastic relaxation or weighted assignment is regarded as more moderate than deterministic assignment. However, it is also commonly recognised that stochastic relaxation, such as simulated annealing, does not guarantee convergence. In fact, the selection of starting temperature or the setting of annealing schedule are often heuristic. An initial temperature, set too high, leads to high computational cost while an initial temperature, set too low, yields similar result as deterministic relaxation but incurs higher computational cost than deterministic relaxation. After intensive testing with stochastic and deterministic relaxation on the data sets we used, we observed that deterministic assignment strikes a better balance between computational cost and clustering accuracy.

## 2.4 Experimental Results

### 2.4.1 Experiment on Simulated Data

Simulated data sets are necessary in evaluating the algorithm performance because the biological meaning of real data sets are very often not clear. Besides, simulated data

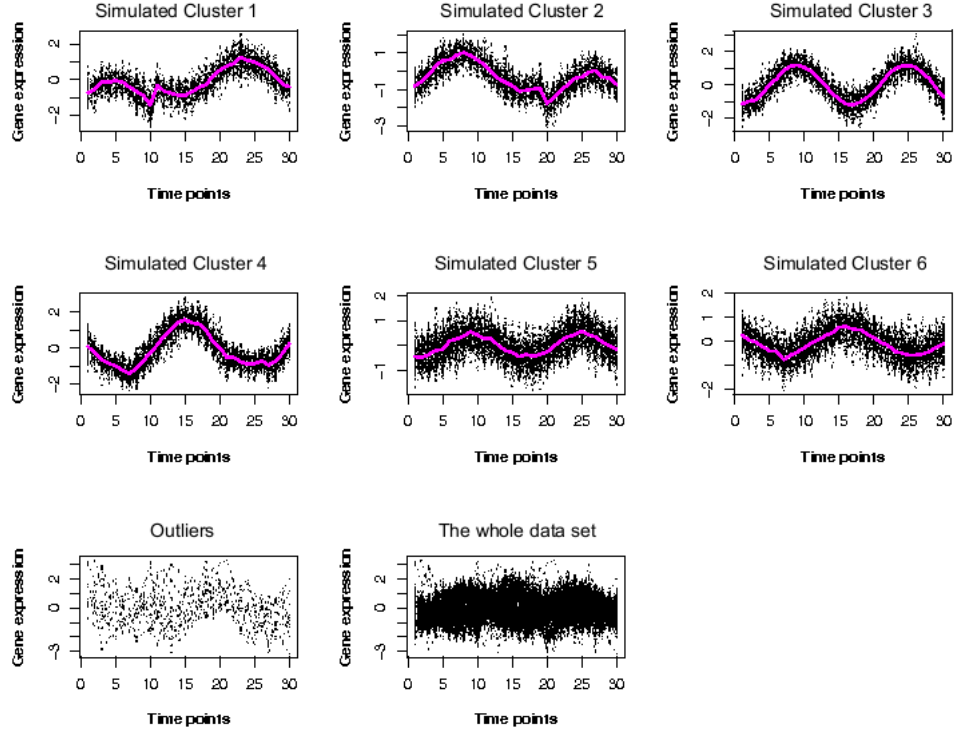


Figure 2.2: The resulting partition by the partial regression clustering algorithm for the simulated data set. The first 6 plots correspond to the gene clusters, the left plot in the third row shows the outliers, the right plot in the last row shows the whole data set.

sets provide more controllable conditions to test an algorithm. However, the simulated data need to share statistical characteristics with biological data.

A simulated data set is generated from a model  $x(i, j) = \alpha_i + \beta_i \psi(i, j) + \varepsilon(i, j)$ , where  $\psi(i, j) = \sin(\gamma_i j + \omega_i)$ .  $\alpha, \beta, \gamma, \omega$  are cluster-specific parameters and are chosen according to a normal distribution with mean equal to 2 and standard deviation 1. This kind of simulation has been used in many studies for clustering validation [97, 116].

In detail, the data set is generated from the following patterns:

$$\begin{aligned}
x1(i, j) &= 0.1 + \sin(1/3j) + \varepsilon(i, j), \\
x2(i, j) &= -0.1 + \sin(1/3j - 1) + \varepsilon(i, j), \\
x3(i, j) &= 1.2\sin(2/5j - 2) + \varepsilon(i, j), \\
x4(i, j) &= 1.5\sin(1/3j - 3.5) + \varepsilon(i, j), \\
x5(i, j) &= 0.5\sin(2/5j - 2.2) + \varepsilon(i, j), \\
x6(i, j) &= 0.6\sin(1/3j - 3.8) + \varepsilon(i, j).
\end{aligned} \tag{2.26}$$

$\psi$  models the cyclic behavior of gene expression patterns. 30 time points are taken from the models in Eq.(2.26), with  $i \in \{1, 2, \dots, 6\}$ ,  $j \in \{1, 2, \dots, 30\}$ . The cluster sizes are 50, 60, 70, 80, 90, 80. To model the noisy environment of microarray experiments, Gaussian noise  $\varepsilon$  is added to all data. In total, 10 outliers are generated by adding large variance Gaussian noise to three sine waves. Altogether, the simulated data set is of size 440. Finally, we manually reduced the amplitude of patterns of two clusters to increase the complexity of the simulated data set. The simulated data in the first two plots in Figure 2.2 have part of their patterns modified and shifted.

The clustering result by the proposed PMDE is depicted in Figure 2.2. The correct partition is achieved, with all ten outliers detected as shown in the seventh plot and the whole data set plotted in the last one.

### 2.4.2 Experiments on Yeast Cell Cycle (Y5) Data Set

A clustering method can be evaluated on theoretical grounds by internal or external validation, or both. For internal validation, a statistical measure is preferred. Our



algorithm is first validated via the CH measure in a comparison with SplineCluster and MCLUST, two of the most popular clustering methods in the literature. On the other hand, a measure of agreement such as the adjusted Rand index (ARI) [64] between the resulting partition and the true partition, if known, is often used as an external validation criterion. Although a lot of evaluations for methods of the same kind are conducted in this way [97, 116, 133, 152], we note that there is currently no ground truth, given our knowledge of the biological structures [39].

Recognising this, we set out to evaluate the performance of our algorithm by systematically finding biologically relevant evidence [43, 73, 107]. The key to interpret a clustering outcome is to recognise the functional relationships among genes within a cluster as well as between clusters. To this aim, we first provide an enrichment analysis for individual clusters based on Gene Ontology [132], one of the most important and widespread ontologies in Bioinformatics. Then, the overall performance between different clustering algorithms is compared by biological validation.

### **Yeast cell cycle (Y5) data set**

The yeast Y5 data set [24] is popular in the clustering literature for its easy accessibility. Expression levels of *Saccharomyces Cerevisiae* measured at 17 time points. A subset of 384 genes are chosen according to their different peak time in five cell cycle phases: Early G1(G1E), late G1(G1L), S, G2 and M [152]. Based on their peak time, this data set was originally clustered into five gene clusters [152], as shown in Figure 2.3 except the bottom right plot. The original partition makes use of only partial information of gene expression which directly leads to the ambiguities between gene clusters. This partly explains why many clustering algorithms have poor performance (with adjusted Rand index [64] lower than 0.5 when it is used as external index [84, 116]). The biological structure is still unclear, even in such heavily investigated

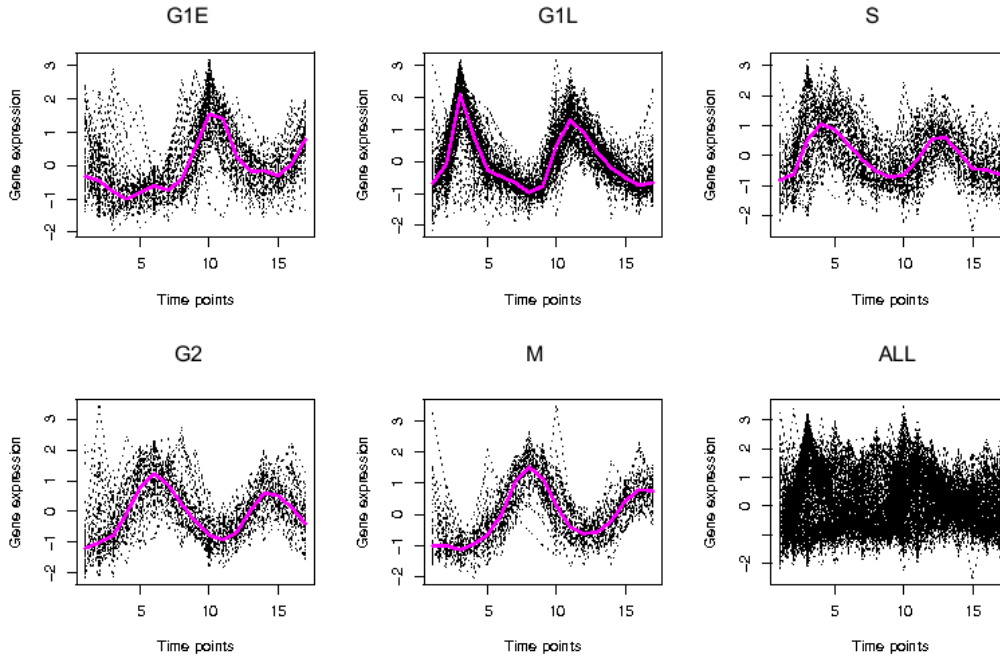


Figure 2.3: The original partition of the yeast Y5 data set with the bottom right plot of the whole data set.

organisms as yeast *Saccharomyces Cerevisiae*.

### 2.4.2.1 Clustering yeast Y5 data set

Table 2.1: Cross tabulation of the original partition and the PMDE clustering partition for the Y5 data set.

	C1	C2	C3	C4	C5	C6	C7	C8	SG	Total
G1E	29	2	12	19	3	0	0	0	2	67
G1L	5	52	0	10	63	4	0	0	1	135
S	1	8	0	2	18	33	11	1	1	75
G2	0	0	0	0	0	7	30	10	5	52
M	1	0	23	0	0	0	1	29	1	55
Total	36	62	35	31	84	44	42	40	10	384

The yeast Y5 data set is chosen not only because it is well-studied in the gene expression clustering literature, but also because of its difficulty in terms of clustering.

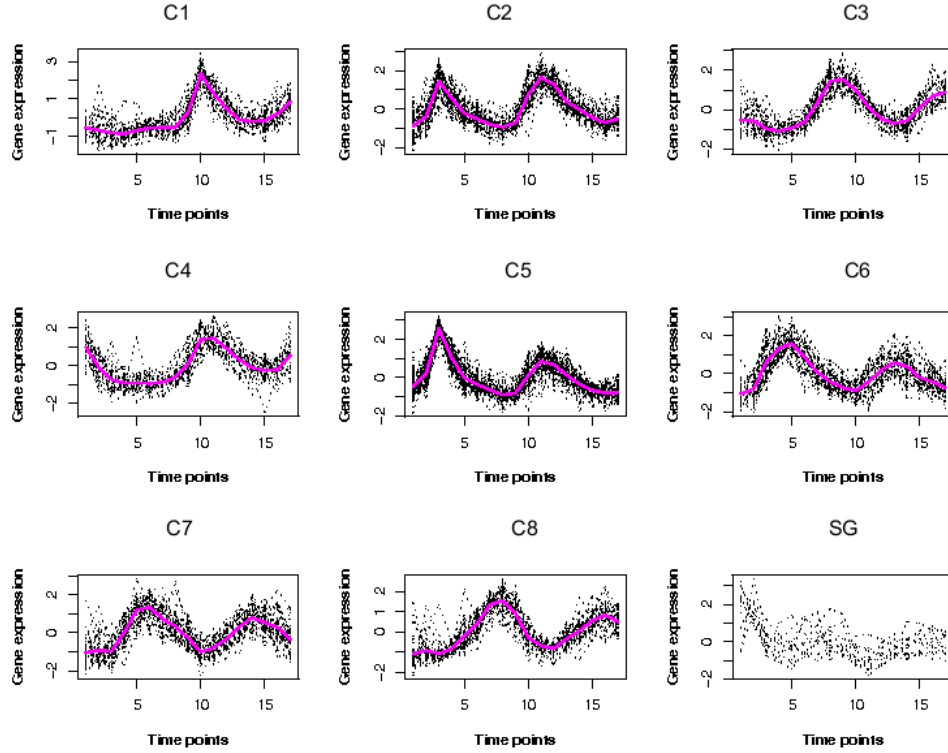


Figure 2.4: The clusters by the partial regression clustering algorithm for the Y5 data set. The bottom right plot shows the scattered genes.

The original partition makes use of only partial information of gene expression which partly explains why many clustering algorithms have poor performance [84, 116]. Moreover, the average cluster size is still far larger than desirable for efficient biological inference, as can be seen from the right-most column of Table 2.1 which contains the size of original partition. It was recently suggested that clustering based on overall profiles is preferred to the original partition on a different subset from the same data set [107]. We employ the proposed partial regression clustering algorithm to partition the Y5 data set into tight clusters. By obtaining tighter clusters, we expect to obtain more informative and efficient biological inference. The tightness threshold  $\nu$  is set to 8 as a result of estimation during the initialisation and the number of knots for the spline

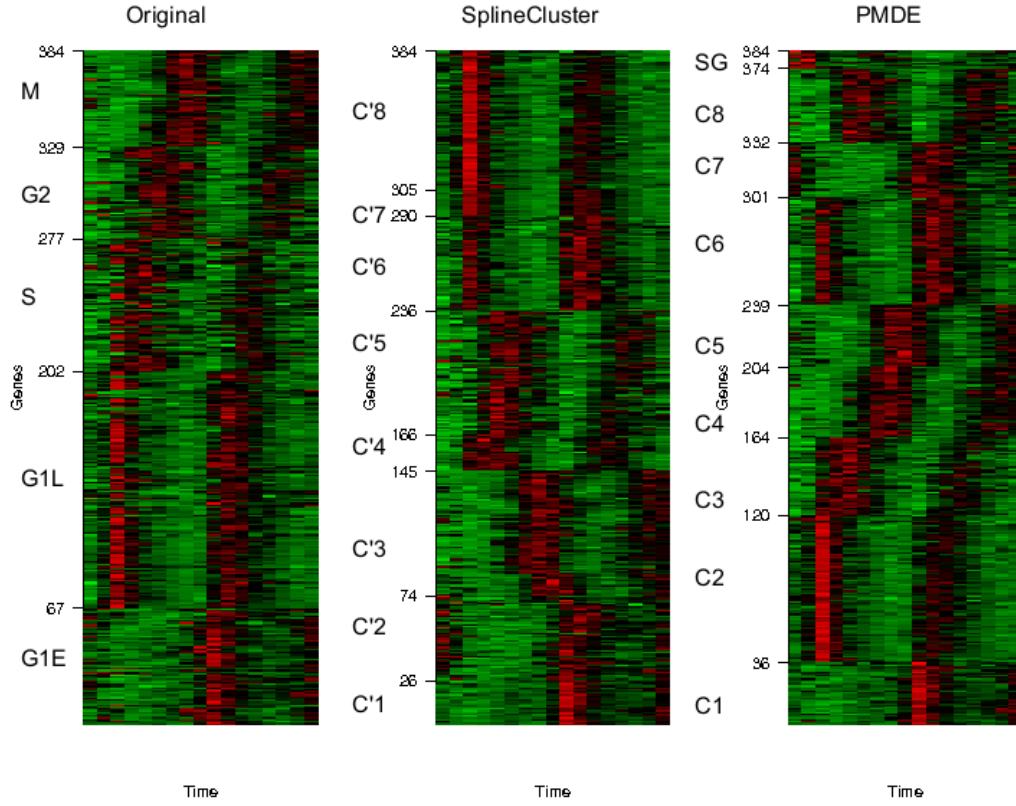


Figure 2.5: Heatmaps for the original partition (left), SplineCluster (middle) partition and the proposed PMDE clustering (right) partition. The brighter red color corresponds to higher expression levels and brighter green color corresponds to lower expression levels.

basis is set experimentally to 13 to allow flexibility of the curve without overfitting.

The clustering outcome of our algorithm is plotted in Figure 2.4. Genes in the bottom right plot are the scattered genes. The eight clusters (C1-C8) with scattered genes (SG) in the new partition are then cross-tabulated with the original partition in Table 2.1. The top row corresponds to the resulting partition, with C1-C8 denoting the eight clusters and SG denoting the set of scattered gene. Each number in the table except the right-most column and bottom row is the number of genes in both clusters corresponding to its row and column. The bottom row indicates the sizes of clusters

of our partition and the right-most column shows those of the original partition. The two partitions agree on many genes but also differ in an interesting way. The new partition reveals neat and easily differentiable patterns. Also, we examined the clustering outcome given by our algorithm and by other algorithms.

First of all, to see the effect of scattered gene detection, three algorithms are compared based on the full data set (384 genes). By controlling a parameter in SplineCluster we obtained 8 clusters for comparison. The partitions of original, SpineCluster and partial regression analysis are illustrated in heatmaps plotted in Figure 2.5 for comparison, where an obvious improvement with respect to class distinction can be seen in the PMDE heatmap. The tick marks on vertical axis in each heatmap indicate where the clusters are located, while in the PMDE heatmap the last (top) cluster corresponds to the scattered genes. The second original cluster (G1L) which is split into the sixth, seventh, and eighth clusters in the SplineCluster partition (C'6, C'7, C'8), and the second and fifth cluster in the PMDE partition (C2, C5). A closer look at the seventh and eighth cluster in the SplineCluster partition shows they differ only slightly in the peak values. However, in microarray data analysis, distinct expression patterns are more interesting than different peak values. This is one of the reasons we use a spline model in our algorithm to capture biologically relevant information. Consider the third cluster in the SplineCluster partition, which is split into the sixth and seventh clusters in our partition. The two clusters show two entirely different patterns, one shifted from the other. Note that a bit cluster corresponding to tick mark 166-236 in the SplineCluster heatmap contains many scattered elements. This is exactly the problem with SplineCluster as stated in Section 2.2.3. From these results, it is obvious that because of its ability in scattered gene detection, our algorithm reveals more distinguishable patterns in the data. The set of scattered genes is listed in Table 2.6 with

their annotations.

Then we use the 374 genes (excluding the 10 scattered genes), and again obtained 8 clusters for SplineCluster. As there is no biological knowledge input, comparison can first be conducted in a purely statistical manner, by the CH index. MCLUST [44] is a widely used mixture model-based clustering method. It is unsupervised, not only in determining the number of clusters, but also in selecting the type of model that best fits the data. The R implementation of MCLUST is used in our experiment. For the 374-gene data set it decided on the EEE (Equal volume, shape and orientation) model and also found 8 components. Our algorithm achieves the highest CH value of 637.4, followed by 588.3 by MCLUST and 523.3 by SplineCluster. Mean CH values for 10 random partitions is 363.3 with standard deviation of 3.23.

### 2.4.2.2 Gene ontology enrichment analysis

To investigate how genes within a cluster are functionally related, and how clustering helps distinguish different functional groups, we apply Gene Ontology enrichment analysis, introduced in Section 1.2.2 to our clustering outcome. In the process, GO terms that are likely to be over-represented in each of the clusters are identified. These GO terms are of interest because they represent the most common functions that the genes in a cluster share.

The probability that a given GO term is over-represented in a gene cluster can be calculated using the hypergeometric distribution [131]. The process proceeds as follows. First, for each cluster, all unique GO terms that are associated with the genes in the cluster are identified. Then for each term, two statistics are needed: the number of genes in the cluster that are annotated to a term and all known genes annotated to a term. With this information, the hypergeometric distribution can be applied to

identify GO terms that are associated to more genes in a cluster than by chance. The probability that a GO term appear not merely by chance is indicated by the resultant  $p$ -values. Using the hypergeometric distribution, suppose there are  $j$  genes annotated to a function in a total of  $G$  genes in the genome, the  $p$ -value of observing  $h$  or more genes in a cluster of size  $b$  annotated to this function is given by [131]

$$p[O \geq h] = 1 - \sum_{i=0}^{h-1} \binom{b}{i} \binom{G-b}{j-i} / \binom{G}{j}, \quad (2.27)$$

where  $O$  is the number of genes annotated with the function. The lower the  $p$ -value is, the more unlikely the null hypothesis that the terms appear by chance is true. In this way, the over-represented terms are found for each cluster.

We analyse the functional categories that are statistically over-represented in the clusters obtained by the proposed algorithms (PMDE clusters) and SplineCluster (SC clusters). For simplicity, we provide the enrichment analysis results in Table 2.2 and Table 2.3 based on the Biological Process Ontology. As indicated by the lowest  $p$ -values in each cluster, all PMDE clusters have a statistically significant set of cell cycle related terms (all lowest  $p < 10^{-5}$ ), while for SC only six out of eight clusters have such significance. We observed that from the remaining two clusters of poorer quality ( $p = 6.35 \times 10^{-3}$  and  $2.51 \times 10^{-4}$ ), some genes involved in DNA replication (*SLD2*, *POL12*, *CDC45* etc. [126]) were combined into PMDE cluster 5, resulting in a tight cluster that has a significantly functional over-representation of DNA strand elongation ( $p = 5.04 \times 10^{-9}$ ) and other functions in DNA replication. Such a high quality cluster is essential for predicting unknown functions of genes such as *YHR151C* and *YNL058C* within the cluster.

Table 2.2: Over-represented GO terms by the proposed PMDE algorithm for the Y5 data set

Cluster	GO ID	GO term	<i>p</i> -values	Gene counts
1	GO:0006118	electron transport	1.06E-06	5
1	GO:0006119	oxidative phosphorylation	5.82E-06	5
1	GO:0042775	ATP synthesis coupled electron transport	1.13E-05	4
2	GO:0006974	response to DNA damage stimulus	1.09E-06	12
2	GO:0045005	maintenance of fidelity during DNA replication	2.56E-06	5
2	GO:0000135	septin checkpoint	3.37E-06	3
3	GO:0006268	DNA unwinding during replication	3.31E-09	5
3	GO:0032392	DNA geometric change	3.49E-08	5
3	GO:0006270	DNA replication initiation	5.54E-07	5
4	GO:0005975	carbohydrate metabolic process	7.61E-06	8
4	GO:0006101	citrate metabolic process	0.000164	2
4	GO:0006091	generation of precursor metabolites and energy	0.000185	7
5	GO:0022616	DNA strand elongation	5.04E-09	8
5	GO:0051276	chromosome organization and biogenesis	1.73E-08	26
5	GO:0009719	response to endogenous stimulus	1.79E-08	17
6	GO:0007020	microtubule nucleation	1.05E-08	6
6	GO:0007017	microtubule-based process	2.92E-08	9
6	GO:0007059	chromosome segregation	1.09E-07	9
7	GO:0000070	mitotic sister chromatid segregation	3.84E-05	5
7	GO:0007001	chromosome organization and biogenesis	4.69E-05	13
7	GO:0016481	negative regulation of transcription	5.08E-05	7
8	GO:0000910	cytokinesis	2.14E-06	7
8	GO:0000278	mitotic cell cycle	1.22E-05	9
8	GO:0000916	cytokinesis, contractile ring contraction	0.000222	2



Table 2.3: Over-represented GO terms by the SplineCluster Algorithm for the Y5 data set

Cluster	GO ID	GO term	<i>p</i> -values	Gene counts
1	GO:0006268	DNA unwinding during replication	7.38E-05	3
1	GO:0006267	pre-replicative complex formation	9.54E-05	3
1	GO:0050790	regulation of catalytic activity	0.000178	4
2	GO:0006260	DNA replication	9.51E-08	10
2	GO:0006310	DNA recombination	9.44E-07	9
2	GO:0006974	response to DNA damage stimulus	9.14E-06	11
3	GO:0022402	cell cycle process	1.63E-06	16
3	GO:0000278	mitotic cell cycle	3.14E-05	11
3	GO:0000074	regulation of progression through cell cycle	3.55E-05	9
4	GO:0022616	DNA strand elongation	1.59E-10	9
4	GO:0006273	lagging strand elongation	5.73E-09	7
4	GO:0006261	DNA-dependent DNA replication	1.35E-07	9
5	GO:0007165	signal transduction	0.006354	4
5	GO:0007154	cell communication	0.010349	4
5	GO:0030541	plasmid partitioning	0.011825	1
6	GO:0009262	deoxyribonucleotide metabolic process	0.000251	2
6	GO:0006259	DNA metabolic process	0.000476	7
6	GO:0006334	nucleosome assembly	0.000587	2
7	GO:0007017	microtubule-based process	9.30E-06	5
7	GO:0007020	microtubule nucleation	4.25E-05	3
7	GO:0009225	nucleotide-sugar metabolic process	9.01E-05	2
8	GO:0007120	axial bud site selection	1.14E-06	5
8	GO:0000819	sister chromatid segregation	1.66E-05	6
8	GO:0000910	cytokinesis	3.97E-05	7

In addition, good agreement was found between known biological functions and gene clusters found by the proposed algorithm. Many clusters in the PMDE partition are significantly enriched with distinctive cell cycle relevant functions, indicating a good separation of functional clusters. For example, cluster 5 has an over-representation of DNA strand elongation ( $P < 10^{-8}$ ), and cluster 6 is enriched with microtubule nucleation and chromosome segregation ( $P < 10^{-7}$ ) which is crucial to chromosome division. Consistent with their biological functions, two clusters involving genes expressed in M and earlier phases reveal patterns of slightly different peak time: cluster 3 contains an over-representation of genes involved in DNA unwinding during replication ( $P < 10^{-8}$ ) and DNA geometric change ( $P < 10^{-7}$ ); and cluster 8 is enriched with cytokinesis that is known to occur after replication and segregation of cellular components. The two gene clusters are both biologically meaningful and statistically sound.

### 2.4.2.3 Predictive accuracy test

We compared five clustering methods: the proposed PMDE algorithm, SplineCluster, MCLUST [44], hierarchical clustering, K-means, in terms of their predictive accuracy [133]. Since the underlying biological ground truth is unknown, evaluation of clustering algorithms for gene data cannot be carried out by similarity measures such as ARI. Instead, predictive accuracy was proposed to test functional prediction accuracy from clustering. The rationale is that since clustering is aimed at functional prediction of novel genes, if a cluster has exceptionally high occurrences of a certain gene annotation  $F$  ( $p$ -value smaller than a certain threshold), all genes in this cluster can be predicted to be in the functional category  $F$ . The ratio of the verified predictions to all prediction made reflects the accuracy of a clustering algorithm. However, we have to

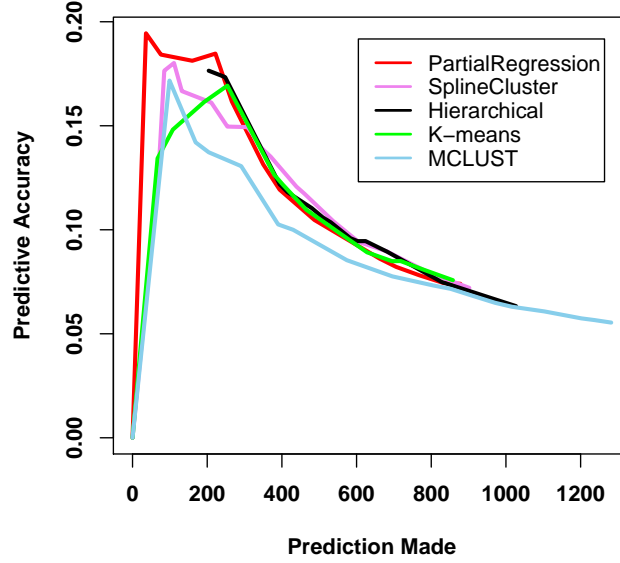


Figure 2.6: Predictive accuracy plots for five clustering methods on Y5 data set. Five clustering methods are evaluated in terms of their functional group prediction accuracy. The five methods are the proposed PMDE (red), SplineCluster (violet), MCLUST (black), hierarchical clustering (green), and K-means (blue). The higher the curve is the better the performance is.

bear in mind that this measure greatly depends on the annotation quality of the data set under study.

Since our results involved a set of scattered genes, we propose as described below a slightly different criterion to the one in [133]. Suppose a functional category,  $F_i$ , has  $v_i$  genes in a data set of size  $n$ . If there are in total  $V$  genes belonging to functional categories  $F_1, F_2, \dots, F_M$ , the remaining  $n - V$  genes are denoted as ‘unannotated’. Such grouping and the resulting partition  $C_1, C_2, \dots, C_K$  of a clustering method can be cross-tabulated to form a table. Let  $n_{ij}$ , ( $i = 1, 2, \dots, M$  and  $j = 1, 2, \dots, K$ ) be the  $(i, j)$  entry of the table denoting the number of annotated genes,  $p_{ij}$  be the corresponding  $p$ -value, and  $n_{.j}$  be the size of cluster  $C_j$ . Given a threshold  $\delta$ , for a  $K$ -cluster solution, its

predictive accuracy  $A$  is defined as

$$A(\delta) = P_V(\delta)/P_C(\delta), \quad (2.28)$$

where  $P_V(\delta)$  is the verified predictions and  $P_C(\delta)$  is the predictions calculated by

$$P_V(\delta) = \sum_{j=1}^K \sum_{i \in \{x | p_{xj} < \delta\}} n_{ij},$$

$$P_C(\delta) = \sum_{j=1}^K \sum_{i \in \{x | p_{xj} < \delta\}} n_{.j}.$$

Table 2.4 lists 68 genes in Y5 data set that are verified to be cell-cycle related to their corresponding cell cycle phases, together with their annotations. Those six cell-cycle related categories plus a group of ‘Not verified’ genes can serve as functional categories, so that seven categories can be cross-tabulated with the new partition as in Table 2.5. The bottom row of Table 2.5 shows the sizes of clusters and the set of scattered genes. All scattered genes are excluded from this evaluation. By pooling results from various thresholds, we obtain curves of ‘prediction made’ versus ‘accuracy’ for all methods in comparison ( $K=8$ ). As shown in Figure 2.6, the curve for the proposed PMDE method is above the others, indicating higher accuracy in functional group prediction.

Table 2.4: Verified cell cycle related (68) genes in the yeast Y5 data set

Cell Cycle	Genes Systematic Names
M/G1 Boundary	YKL185W YLR274W YBR202W YJL194W YAL040C YLR286C YDL127W YDL179W YGR044C YLR079W YER111C YBR083W
Late G1, SCB regulated	YMR199W YPL256C YDL227C YER001W YNL289W YJL187C YBR067C
Late G1, MCB regulated	YJL115W YDL197C YOR074C YLR103C YDL164C YPR120C YGR109C YPR175W YBR278W YDR309C YDL003W YOL090W YDR097C YKL101W YDR113C YNL082W YNL102W YBL035C YNL262W YBR088C YKL045W YKL113C YAR007C YNL312W YJL173C YER070W YDR356W YKL042W YPL153C YJL092W YML021C
S-phase	YBL003C YBL002W
S/G2-phase	YMR198W YPR141C
G2/M-phase	YLR131C YOR058C YGL116W YMR001C YGR108W YPR119W YGR092W YJL157C YAR018C YIL106W YBR054W YDR033W YHR152W YDR146C

Table 2.5: Cross-tabulation of clustering outcome (C1-C8 and SG) with verified gene functional categories for the yeast Y5 data set

Cell Cycle	C1	C2	C3	C4	C5	C6	C7	C8	SG	Total
M/G1 Boundary	7	2	1	2	0	0	0	0	0	12
Late G1, SCB regulated	0	2	0	0	3	1	0	0	1	7
Late G1, MCB regulated	0	13	0	0	15	3	0	0	0	31
S-phase	0	1	0	0	0	1	0	0	0	2
S/G2-phase	0	0	0	0	0	1	1	0	0	2
G2/M-phase	2	0	3	1	0	0	0	8	0	14
Not verified	27	45	32	28	66	37	41	33	9	316
Total	36	62	35	31	84	44	42	40	10	384

## 2.4.2.4 Scattered genes

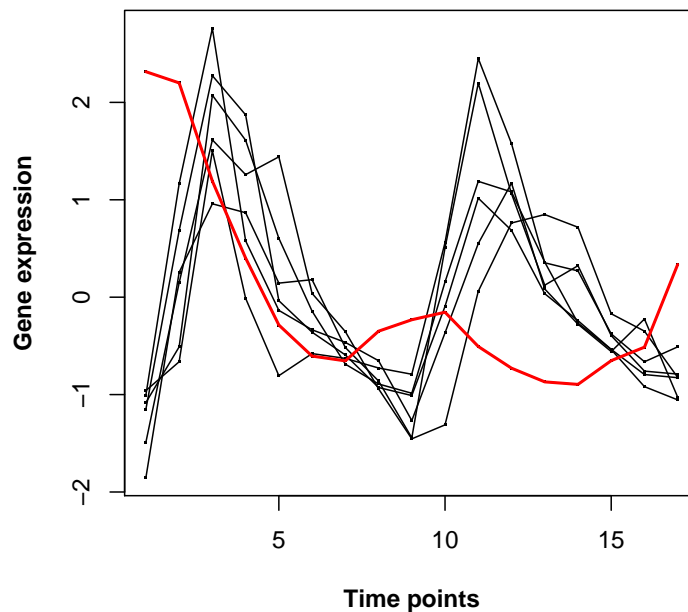


Figure 2.7: The profiles of seven genes related to Late G1, SCB regulated cell cycle phase. The red profile is the gene “TIP1/YBR067C”, one of the ten scattered genes. It displays a distinctive pattern from the other six genes annotated to be in the same functional group.

Another important aspect in our investigation is to study the set of scattered genes. Multiple experiments are conducted with various tightness thresholds,  $\nu$ , in our partial regression method. In Table 2.6, the set of scattered genes found in eight runs of our program with various thresholds and their annotations are listed. Their frequencies of appearance in these experiments are shown in the column Freq. (out of 8). We noticed that although these thresholds result in different numbers of clusters, the set of scattered genes hardly changes (Table 2.6, column Freq.). Such consistency leads one to think about the underlying biological meaning. As has already been pointed

out [71], scattered genes can be those individuals that are not relevant to the biological process under study. However, we stress here that they can also be of significant interest, as each of them might be a key component of the cell cycle that may affect other components and indeed may be a transcription factor themselves. Therefore, its expression pattern can be uncorrelated to others in the set under study. Alternatively, a scattered gene can represent a gene whose expression is controlled by more transcription factors than the other co-regulated genes within clusters. Moreover, because the set of genes under investigation is usually selected after performing gene ranking, there may be others in the complete list that would cluster with scattered genes. All these considerations drove us to further investigate this set of scattered genes.

Among the scattered genes, five are either not well-understood or unknown for their functions. Only one of them, *TIP1/YBR067C*, is verified to be cell cycle related in phase Late G1, SCB regulated (Table 2.4, second group). Indeed, according to Table 2.4, one would conclude that all the seven genes in Late G1, SCB regulated phase to have the same behaviour. However, when their profiles are plotted as in Figure 2.7, we can see that *TIP1/YBR067C* is uncorrelated to the others, making it an interesting subject for further study.



Table 2.6: Details of the set of scattered genes for the yeast Y5 data set detected by PMDE, including their SGD IDs, the frequencies that they are found across eight experiments of various thresholds, and their annotations.

Gene	SGD ID	Freq.	Function
BRE2/YLR015W	SGD:S000004005	7/8	Subunit of the COMPASS (Set1C) complex, which methylates histone H3 on lysine 4 and is required in transcriptional silencing near telomeres
MOD5/YOR274W	SGD:S000005800	7/8	Delta 2-isopentenyl pyrophosphate:tRNA isopentenyl transferase, required for biosynthesis of the modified base isopentenyladenosine in mitochondrial and cytoplasmic tRNAs
PPR1/YLR014C	SGD:S000004004	7/8	Zinc finger transcription factor containing a Zn(2)-Cys(6) binuclear cluster domain, positively regulates transcription of genes involved in uracil biosynthesis
RNP1/YLL046C, YNL016W	SGD:S000003969	8/8	Ribonucleoprotein that contains two RNA recognition motifs
TIP1/YBR067C	SGD:S000000271	8/8	Major cell wall mannoprotein with possible lipase activity
UIP4/ YPL186C	SGD:S000006107	8/8	Protein of unknown function that interacts with Ulp1p, a Ubl (ubiquitin-like protein)-specific protease for Smt3p protein conjugates
YBR184W	SGD:S000000388	8/8	Putative protein of unknown function
YDL124W	SGD:S000002282	8/8	NADPH-dependent alpha-keto amide reductase
YDR366C	SGD:S000002774	5/8	Hypothetical protein
YLL047W	SGD: S000003970	8/8	(Not annotated)

## 2.4.2.5 Comparative evaluation on scattered gene detection

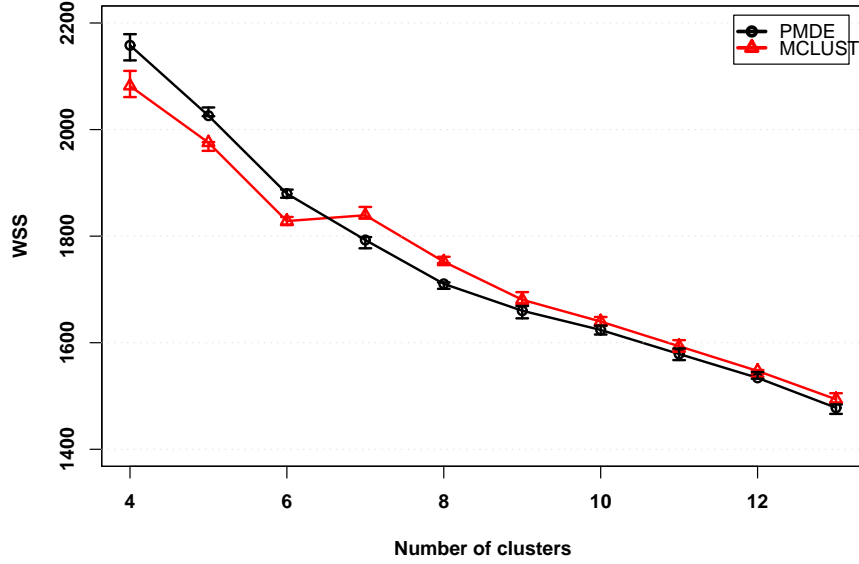


Figure 2.8: Comparison of performance of PMDE and MCLUST in outlier detection. A small index value of  $WSS$  indicates better performance in outlier filtering. PMDE performs better than MCLUST with large number of clusters.

To further assess the proposed PMDE's strength of scattered gene detection, the proposed algorithm is compared with a recent modification of the MCLUST, which allows an additional component of homogeneous Poisson process for scattered genes/noise [47]. The idea is for each method to filter out scattered genes and then, instead of analysing the scattered genes, compare the quality of the filtered data sets in terms of within-cluster sum of squares  $WSS$  as defined in Eq.(2.25). If an algorithm is stronger in outlier filtering, tighter clusters should be found in the filtered data set, hence a smaller value of  $WSS$ . Since the number of scattered genes identified by the two methods may vary, when the sets of scattered genes filtered out by different methods are of different sizes, we randomly sample a subset of the same size as the smaller set

from the larger one and return the leftovers to the filtered data set so that the filtered data sets to be investigated/clustered are of the same size. Because the clustering quality may be affected by the returned genes, we repeat the process of the random sampling of scattered genes and the clustering of the filtered data set 10 times, and take the average value of  $WSS$  to compare against the  $WSS$  of the clustering result by the other method.

We obtain clustering results with the number of clusters  $K$  ranging from 4 to 13 for Y5 data set from both the PMDE and the MCLUST. The results are plotted in Figure 2.8. We can see that the proposed PMDE performs better with large number of clusters,  $K$ , but not as good as the MCLUST with smaller  $K$ . However, this does not mean that the MCLUST outperforms the PMDE because the PMDE is designed to start with an initial set of clusters and iteratively split the current clusters if the splitting can lead to tighter clusters. Therefore, the clustering results by the PMDE with smaller values of  $K$  are not “final” but just “provisional”; when compared to the “final” results by the MCLUST, the performance of the PMDE appears to be inferior. However, when the results by the PMDE is more mature as  $K$  gets bigger, for example when  $K$  is greater than or equal to 7 as shown in Figure 2.8, the proposed PMDE consistently outperforms MCLUST.

### 2.4.3 Experiments on Yeast Galactose Data

#### Yeast galactose data set

The yeast galactose data set [67] consists of gene expression measurements during galactose utilization in *Saccharomyces cerevisiae*. Expression levels were measured across 20 experimental conditions representing 20 perturbations in the GAL pathway.

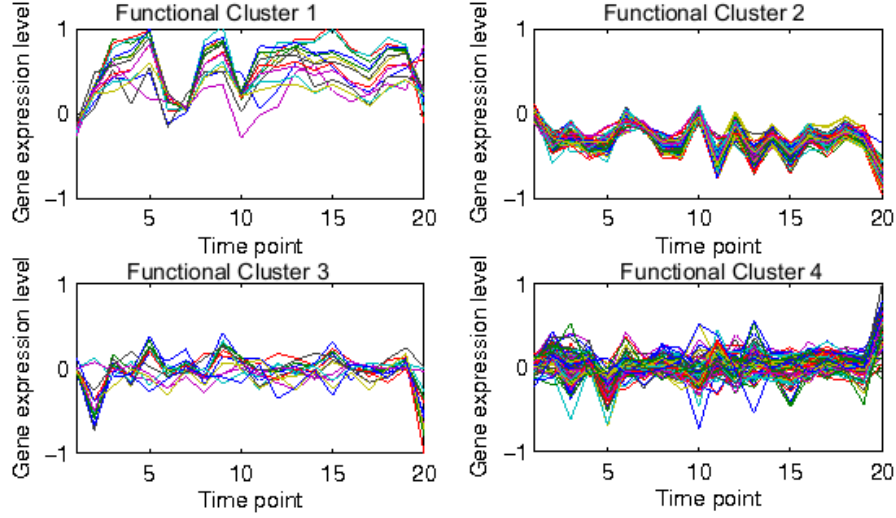


Figure 2.9: Expression data across 20 time points in four functional categories of yeast galactose data.

A subset of measurements of 205 genes whose expression patterns reflect four functional categories in the GAL pathway was chosen and clustered previously [97, 134]. The four gene categories given as ground truth reflect four functional categories. Compared with Y5 data set, yeast galactose data set show more distinguishable expression patterns, as can be seen from Figure 2.9. This data set can represent a case when the experimental data are agreeable to existing functional interpretations [134].

Experiments are conducted on the yeast galactose data set, which has more agreeable correlations to its functional interpretation than the yeast Y5 data. For this data set, our partial regression algorithm yields 4 clusters and 4 scattered genes when the tightness threshold is set to low value. The four clusters (C1-C4) with scattered genes (SG) are then cross-tabulated with the original partition in Table 2.7. We take 4 as cluster number, since it is also in accordance with prior knowledge, and get partitions from all five algorithms. The bottom row of Table 2.7 contains cluster sizes for the original

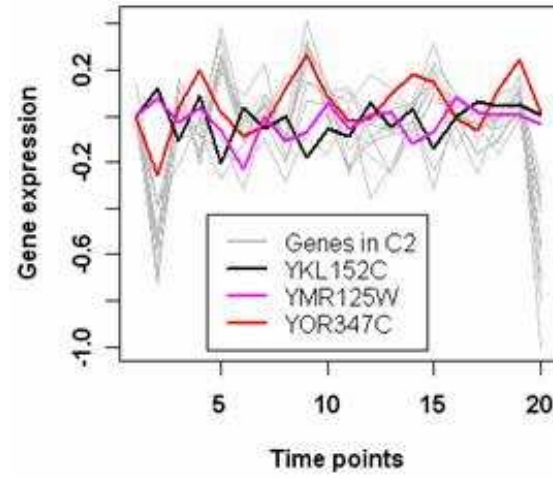


Figure 2.10: Scattered genes in original cluster 2 of the yeast Galactose data set. The expression profiles of some scattered genes detected by the proposed algorithm are plotted for the yeast Galactose data set. This plot shows the expression patterns of all 15 genes in original cluster 2, among them the 3 colored genes are the detected scattered genes.

partition and the right-most column contains cluster sizes for the resulting partition. Each number in the table except the right-most column and bottom row is the number of overlapping genes in both clusters corresponding to its row and column. As a mean of statistical validation, CH measure is applied to the above five algorithms PMDE, Spline Cluster, Hierarchical, K-means, and MCLUST, respectively, giving values of 365.6, 331.1, 360.1, 255.3, and 364.5, respectively. Since there is no given functional categories for this data set, the predictive accuracy index cannot be applied. Instead, we focus on evaluating the power of PMDE in scattered gene detection.

There are interesting findings from the investigation of the set of scattered genes. For instance, one gene (*YMR125W*) belonging to the original cluster O2 is classified as a scattered gene. Of the other 14 genes in original cluster 2, 12 are clustered into C2, 1

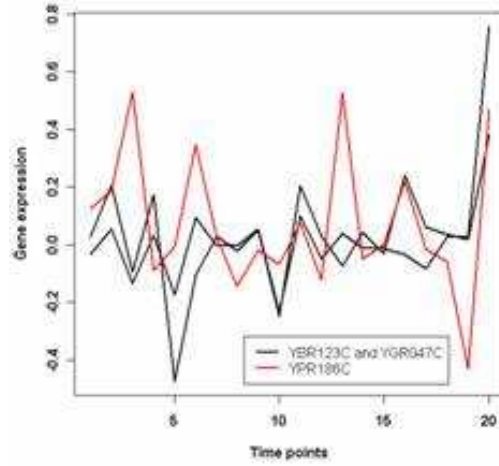


Figure 2.11: Scattered genes in original cluster 3 of the yeast Galactose data set. The expression profiles of the 3 scattered genes in original cluster 3. They share GO annotations but have various expression patterns.

Table 2.7: Cross-tabulation of the original partition (O1-O4) and the resulting partition (C1-C4 and SG) for the yeast Galactose data set.

Cluster	O1	O2	O3	O4	Total
C1	83	0	0	0	83
C2	0	12	0	0	12
C3	0	1	90	1	92
C4	0	1	0	13	14
SG	0	1	3	0	4
Total	83	15	93	14	205

in C3 (*YKL152C*) and 1 in C4 (*YOR347C*). The expression data of all of the 15 genes are plotted in Figure 2.10, revealing very different expression patterns of the 12 genes and the 3 genes differentiated by our algorithm. Both *YKL152C* and *YMR125W* are up-regulated at the beginning with down regulations for all others. The resulting cluster C2 by partial regression is verified by GO, since the 12 genes share similar annotations among the 15 genes in the original cluster O2, for example they are all annotated to Glycolysis (GO:0006096) observed from the Table 2.8.

Table 2.8: Over-represented terms in each original cluster for the yeast galactose data set.

Cluster	GO ID	GO term	<i>p</i> -values	Gene counts
1	GO:0006412	translation	4.37E-95	83
1	GO:0044249	cellular biosynthetic process	1.46E-64	80
1	GO:0044260	cellular macromolecule metabolic process	4.96E-52	80
1	GO:0019538	protein metabolic process	5.69E-52	80
1	GO:0008152	metabolic process	4.85E-24	83
2	GO:0006096	glycolysis	9.53E-29	12
2	GO:0019320	hexose catabolic process	1.22E-25	12
2	GO:0046164	alcohol catabolic process	2.24E-24	12
2	GO:0044275	cellular carbohydrate catabolic process	1.02E-22	12
2	GO:0006094	gluconeogenesis	3.06E-16	8
3	GO:0043170	macromolecule metabolic process	1.53E-35	92
3	GO:0044238	primary metabolic process	3.99E-31	93
3	GO:0044237	cellular metabolic process	6.52E-28	93
3	GO:0000398	nuclear mRNA splicing, via spliceosome	9.31E-28	27
3	GO:0000375	RNA splicing, via transesterification reactions	1.07E-26	27
4	GO:0008643	carbohydrate transport	2.45E-26	12
4	GO:0008645	hexose transport	4.39E-25	11
4	GO:0051234	establishment of localization	7.64E-10	13
4	GO:0015766	disaccharide transport	0.002408395	1
4	GO:0015771	trehalose transport	0.002408395	1

As an important transcription factor, *YPR186C* is an essential protein that binds the 5S rRNA gene through the zinc finger domain and directs assembly of a multi-protein initiation complex for RNA polymerase III. Belonging to the original cluster O3, *YPR186C* is classified as a scattered gene. We plot its expression levels together with two other genes that are also annotated to GO:0006384 (transcription initiation from RNA polymerase III promoter), and found differences among their patterns in Figure 2.11. Since this term is quite specific and it should be able to reflect a gene's function, mechanisms behind such diverse behaviours are still unclear and are worth further investigations.

## 2.5 Conclusions

The aim of clustering gene profiles is to find possible functional relationships among thousands of genes on a microarray. As microarray technique advances, current clustering methods are no longer adequate for some tasks. In response to some of the recent issues, we proposed in this chapter a PMDE algorithm for tight clustering gene expression data. The tightness of resulting clusters can be controlled by a threshold which in a sense decides the number of clusters.

The contributions of this chapter include introducing MDE and the idea of partial modelling to gene expression research, giving comparison of MDE with the most common estimator in the literature - maximum likelihood, and proposing a novel partial regression clustering algorithm. The proposed algorithm can be applied over an existing clustering to get tighter and therefore more informative clusters. In summary, the proposed system benefits from

- the robustness of minimum distance estimator (MDE) to detect scattered genes,
- the idea of partial modelling for obtaining tight clusters,
- the spline regression model for capturing the expression curves at either uniformly or unevenly distributed time points.

In particular, we propose that while the model for data fitting should be sensitive enough for discriminating individuals/genes, the parameter estimator should be robust enough against noise and possible outliers. Therefore, we focused on the differences between estimators by providing experimental comparisons. The robustness of the MDE makes it stand out in our study. An immediate advantage is that when applied to gene expression clustering, it is capable of locating the key components in an unsuper-



vised manner. As a result, a set of scattered genes that has low correlations is naturally obtained.

Although PMDE demonstrates its effectiveness through the comparison with the maximum likelihood method, it also has its limits such as relative inefficiency. The aim of this chapter is not to prove which one is better, but rather to provide analytical examples, discussions and insights for further research.

During the evaluation of the clustering algorithm, we feel, that although GO provide a wealth of complementary biological knowledge that has been cumulated over time, there is currently no best way to utilise it for clustering validation. Indices such as the predictive accuracy abound [29, 133]. However, they take as input GO terms to be used as functional categories. This is problematic, since the uneven granularity and variability of relevance in the GO structure result in that GO terms cannot be compared on the same level. A validity index specifically designed for GO is therefore needed in order to make precise inference. In the next chapter, we propose a new GO validity index designed for this purpose.

## **Chapter 3**

# **Quantitative Assessment of Clustering Based on Gene Ontology**

### **3.1 Introduction**

As the initial step towards biological inference from microarray gene expression data, clustering is crucial in reducing redundant information and identifying key components in the data, as described in Chapter 2. With the prevalence of various clustering algorithms, it is non-trivial to select one that can best tackle the challenges in the data set under study. On the other hand, Gene Ontology (GO) provides a wealth of complementary biological knowledge, which, if properly exploited, can be of great help in assessing clustering algorithms. However, varying levels of biological specificity of curated information and the graph structure of GO hinder quantitative access. Systematic formulation is therefore needed for biological validation of clustering methods.

To this aim, we design specifically for GO a clustering validation index, which consists of two indices measuring the within-cluster functional compactness and the

between-cluster functional similarity, respectively. This chapter is organised as follows. In this section, we give an introduction to GO and the proposal of GO-based clustering validation, providing analytic reasoning to support the proposition. Section 3.2 and 3.3 reviews research trends in GO clustering validation, bringing up prevailing challenges. For evaluation purpose, existing GO-driven validation methods are categorised into two main sets: methods that use GO terms as functional categories and methods that are based on previously defined semantic similarity measures. We empirically prove that the methods in the second category may not be suitable in Section 3.3. Later in Section 3.5, some methods in the first category will be compared with our method proposed in Section 3.4.

Ideally, a validation method should be robust against the noise in GO, and computationally efficient enough to facilitate comparison between different clustering algorithms. It should also take into account not only the sets of GO terms annotated to the gene clusters, but also their significance to the clusters and their specificities to the whole GO structure. Clustering validation techniques based on GO annotation should therefore incorporate both a robust infrastructure and an effective representation of relationships between GO terms. So far, there have been numerous works dedicated to statistical validation of gene expression clustering (see [13] for a good review). However, less attention has been paid to objective clustering validation considering these needs. Moreover, little systematic evaluation on the robustness and effectiveness of various GO-driven validation methods has been performed.

In this chapter, systematic evaluations, including comparison of various clustering algorithms, perturbation experiment, and test on finding optimum cluster number, are provided in Section 3.5 to prove the suitability of the proposed index. Evaluation is performed based on the applications of six popular clustering algorithms to three

biological data sets of diverse features, including two *Saccharomyces cerevisiae* data sets and a *Arabidopsis L. Heynth* data set. In addition, five of the existing GO-driven and data-driven validity indices are used for comparison, providing useful insights on the validity indices and the clustering methods. Excellent performance is observed for the proposed validation index throughout all experiments. While existing methods tend to ignore the redundant and complex features of GO, the proposed index proves to be useful tools for handling these features.

### 3.1.1 An Introduction to Gene Ontology (GO)

As one of the most important and widespread ontologies in Bioinformatics, Gene Ontology (GO) is a structured vocabulary intended for annotating gene products with a consistent, controlled and structured vocabulary. Over the years, GO has become one of the most comprehensive man-curated collections of biological knowledge. For example, 67.4% of yeast gene products (4232/6275, including verified and uncharacterised ORFs, transposable element genes, and all RNA gene products, as of Oct. 2008 [23]) are annotated by one of the three GO categories, the biological process.

The three GO categories in GO are biological process (BP), molecular function (MF) and cellular component (CC), each structured as a directed acyclic graph with nodes representing the GO terms and directed edges representing parent-child relationships between terms. An example of such GO structure in the category BP (GO:0008150) can be seen in Figure 3.1. A directed edge indicates either the child node/term is a subclass (*is\_a*) or a component of the parent node/term (*part\_of*). A term and all its children in the hierarchy can be viewed as a functional cluster. Therefore, in addition to describing the relationships between terms, GO helps set up a two-way



is not necessarily as biologically specific as the other terms at the same level.

Ideally, a good clustering algorithm should produce gene clusters with non-overlapping functions. However, even a perfect partition cannot achieve non-overlapping GO functional annotations, because of the existence of general GO terms. Such overlapping annotations incurred by general terms, if not properly dealt with, will introduce ambiguities in clustering validation. A clear boundary should be drawn between overlapping annotations incurred by general terms and by the fault of clustering algorithm itself. An example of functional overlapping in gene clusters is illustrated in Figure 3.2, where the relationships between GO terms and gene clusters are clearly shown. In this example, the notion “over-represented terms” refers to GO terms that can represent relevant functions of gene clusters, as selected on the basis of their specificities. In this sense, the overlapping GO term for gene clusters C1 and C2 in Figure 3.2 is specific enough to indicate inability of the clustering algorithm.

### 3.1.2 Rationales for GO-based Clustering Validation

A concern about clustering validation based on current biological knowledge, is the often observed contradictions between machine learning results from experimental data and the existing annotations. Clustering identifies groups of genes involved in co-regulated biological processes, or groups that encode functionally related proteins for specific pathways. However, the assignment of a gene to a certain cluster based on its expression and genetic co-regulation based on current knowledge in transcriptomics do not necessarily coincide. Genes known to be involved in a common pathway can end up in completely different clusters, while genes with different functions can be assigned to the same cluster.

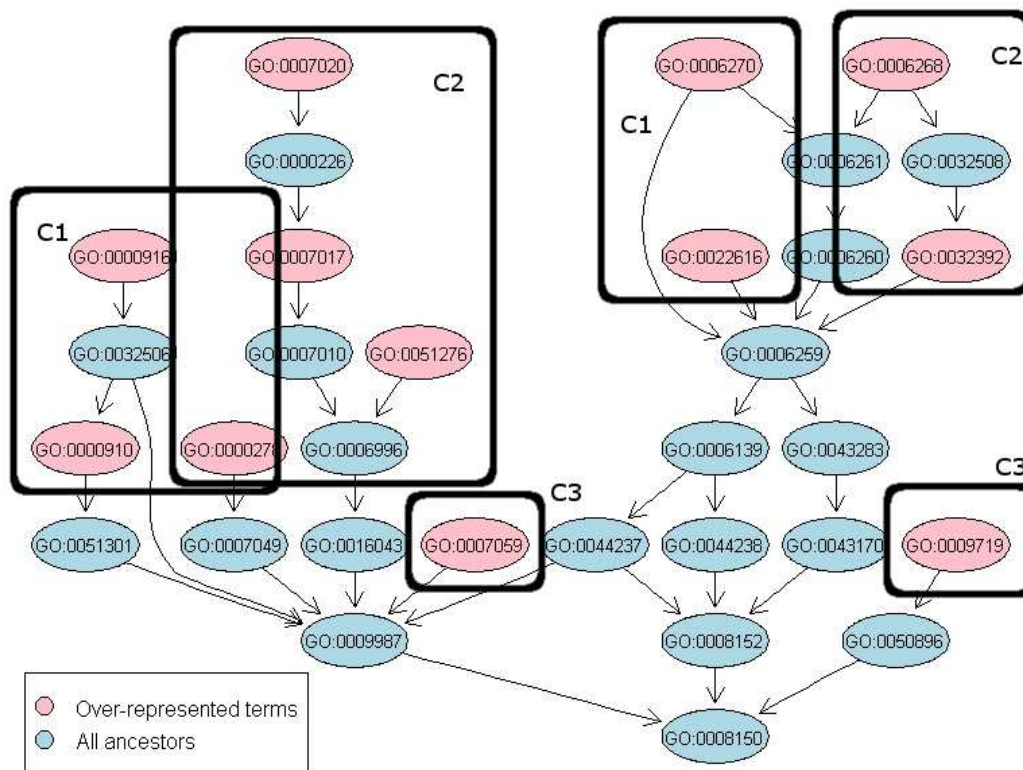


Figure 3.2: An example of functional overlapping in gene clusters with the over-represented terms (pink) for three gene clusters (C1, C2 and C3). There is an overlapping over-represented term (GO:0000278) between C1 and C2.

The reasons of the above contradictions are manifold. First, due to the limited knowledge in annotation, some underlying regulations may be unknown. Existing annotations, however, are skewed towards processes of popular interests [99]. Another reason lies in the microarray data and the clustering algorithm itself. If the clustering algorithm is sensitive to the statistical variation and noise bound in the experimental data, the clustering outcome is less likely to conform to the functional groups. Other reasons lie mainly in the biological responses. For example, cellular processes are affected by both up- and down- regulations and many processes are only regulated by post-translational modifications. Hence, it is possible that gene functions are not

captured by the corresponding expression levels.

Consequently, contradictions between statistical learning and current biological knowledge motivate researchers to explain and uncover the underlying mechanism. They also make the validation of clustering methods an interesting and challenging issue. The interesting aspect is that such contradiction between expression data and functional annotations may ultimately suggest new associations and pathways. Even simple pairwise comparisons can reveal novel interactions in the validation process [20]. The accompanying challenge is the difficulty to design a quantitative evaluation based on biological knowledge, with the existence of annotation gaps. For the purpose of gene function discovery, it is therefore preferable to obtain clusters from purely data-driven methods and evaluate the clusters with existing biological knowledge. This not only prevents clustering results from being biased to current knowledge, but also entails objective validation based on annotations such as GO.

## 3.2 Existing Methods Assuming GO as Functional Categories

Recently, a number of functional validity indices are applied to gene clustering validation using GO terms as functional categories. These methods assume there are known functional categories for at least a subset of genes and assess cluster quality based on the cluster assignments of these genes. Examples are predictive accuracy by Thalamuthu *et al.* [133], entropy-based metrics [89], biological homogeneity index (BHI) and biological stability index (BSI) by Datta and Datta [30].

The measure of predictive accuracy was introduced before in Section 2.4.2.3. The



two entropy-based metrics [89] were proposed to measure the behavioural homogeneity within a cluster and the maximum separation of behaviour across clusters, by strictly mapping genes to functional behavioural groups defined by GO terms. However, as discussed previously, these GO terms are not necessarily comparable with regard to their biological specificities. Also, such simplification of annotations, without taking the GO structure into account, limits the intake of information provided by GO. In this chapter, we select BHI and BSI as representative cases and analyse their effectiveness in the comparative experiments in Section 3.5.

BHI measures how biologically homogeneous the gene clusters are. Intuitively, the measure examines whether the genes placed in the same statistical cluster also belong to the same functional classes. Consider two annotated genes  $i, j$  that belong to the same statistical cluster  $C_k$  in a partition  $P$ ,  $P = \{C_k | k = 1, 2, \dots, K\}$ . Let  $f(i)$  denote the functional class/classes containing gene  $i$  and  $N_k$  denote the number of annotated genes in clusters  $C_k$ . BHI for partition  $P$  is defined as

$$BHI(P) = \frac{1}{K} \sum_{k=1}^K \frac{1}{N_k(N_k - 1)} \sum_{i \neq j \in C_k} \mathbf{I}(f(i) = f(j)), \quad (3.1)$$

where the indicator function  $\mathbf{I}(f(i) = f(j))$  is assigned value 1 if  $f(i)$  and  $f(j)$  match and value 0 otherwise. In the case of multiple functional class assignments for the same genes, any one match is sufficient. If these functions have different relevance, however, the judgement by the indicator function may not be indicative of the real biological meaning.

BSI inspects the stability of clustering for genes with similar biological functions. Each time a sample/time point is removed from the gene expression time series data, and the cluster membership for genes with similar functional annotation is compared

### 3.2 Existing Methods Assuming GO as Functional Categories

---

with the cluster membership using all available samples. Let  $C^{i,x}$  denote the cluster containing gene  $i$  in the clustering based on the reduced expression profile without the  $x$ th sample, and  $C^{i,0}$  be the cluster containing gene  $i$  using the full expression profile,

$$BSI(P) = \frac{1}{F} \sum_{k=1}^F \frac{1}{N_k(N_k - 1)m} \sum_{x=1}^m \sum_{i \neq j \in f_k} \frac{N(C^{i,0} \cap C^{j,x})}{N(C^{i,0})}, \quad (3.2)$$

where  $F$  is the total number of functional classes,  $m$  the number of samples/time points, and  $N(\cdot)$  denote size or cardinality. This measure is based on the tenet that a stable clustering algorithm would produce similar answers, as judged biologically, based on the full and the reduced data. Thus, the clusters using full and reduced data containing two functionally similar genes should have substantial overlaps. Since the index examines whether the cluster membership for genes with similar functional annotation remain the same when a sample is removed, accuracy of this index may largely depend on the quality of data.

Values for both of BHI and BSI are bounded by  $[0, 1]$ , with larger scores of BHI corresponding to more biologically homogeneous clusters, and larger scores of BSI corresponding to more stable clusters of the functionally annotated genes, respectively. Since there is no concept of depth for GO, it is difficult to find GO terms that have the same biological specification and relevance. To apply both indices to GO validation, a threshold is used to select biologically specific GO terms as functional classes. Consequently, all selected terms are treated on the same level. In fact, if such functional categories are known even only for a subset of genes, one can always assign the rest of the genes to one category and then the widely-used Adjusted Rand Index [64] can be utilised for assessing the performance of clustering algorithms. However, when GO categories with uneven level of biological relevance are used, such assessments are

not conducted on the fair ground. To prove this point, we will compare our proposed validation measure with BHI and BSI in Section 3.5.

## 3.3 Existing Methods Based on GO Semantic Similarity

### 3.3.1 GO Semantic Similarity

Currently, the majority of GO-driven clustering validity indices heavily relies on semantic similarity measures for GO terms (term-term similarity) [69, 85, 109]. Based on these measures, pairwise relationships of gene products can be set up by mapping GO terms to genes and thereby enable distances among gene clusters to be mapped out. Next, we briefly review some of these techniques in a hierarchical style.

#### **Term-term similarity**

Semantic similarity measures for GO terms often take into account the information content of GO terms [109] and GO's graph structure. Information content is a useful criterion indicating the usage frequency of a term. The assumption is that the less frequently a term occurs, the more informative it is since it is more specific. Although this assumption is not always true [99], information content serves as a practical guidance to the specificity of a term if no other information is available. One of the most popular term-term similarity measure, Resnik's measure [109] is defined as the information content of the lowest common ancestor of the two terms. Following, a number of measures were proposed as improved versions of Resnik's measure. For instance, Lin's similarity measure [85] and Jiang and Conrath's distance measure [69] take into account the information content of both two terms and their lowest common ancestor,

but differ in the way of normalisation. Relevance similarity [115] was proposed as Lin's measure with weight assignments, signifying a favour for biologically relevant terms in the comparisons.

#### **Gene-gene similarity**

Based on the above term-term similarities, a most common measure for the similarity between two gene products, each mapped to a set of GO terms, is often calculated as the average or maximum pairwise similarities between the two sets of terms [144]. Other methods have been proposed. FuSSiMeg enriched term similarity by Couto *et al.* [26] takes the maximum term-term semantic similarity measure times the corresponding information content for both terms, in order to take into account the significance of a term. FunSim score [115] makes use of a similarity matrix whose elements are the pairwise similarities between terms. The score is taken as the average over the row maxima or column maxima, whichever is higher. The final score is computed by averaging scores based on ontology MF and BP, respectively.

#### **Cluster-cluster similarity**

With the availability of gene-gene similarity, cluster-cluster similarity can be defined to assess clustering quality. This is traditionally achieved with existing data-driven validation indices by simply replacing the similarity measure with one of the above gene-gene similarity measures [15, 125]. For example, Bolshakova *et al.* [15] used C-index and Goodman-Kruska index with Wu and Palmer's semantic measure [103] and Resnik's measure for clustering validation and optimal cluster number selection.

#### 3.3.2 Problems of Methods in this Category

To assess semantic similarity measures for term-term relationship, many studies quantitatively correlate the semantic similarity measures with various genomic features [14, 57, 90]. Established approaches use genomic features such as sequence, expression and interactions to define gene-gene similarity. The assumption is that a good agreement between such similarity and gene-gene semantic similarity may suggest a good semantic measure.

Another assumption is that highly similar sequences should be highly semantically similar. Lord *et. al* [90] were among the first to compare Resnik, Lin's, Jiang and Conrath's measures by correlating the average semantic similarity with protein sequence similarity using BLAST's [1] bits score. Although none of the three measures has clear advantage over the others, Resnik's measure shows highest correlation between sequence similarity and semantic similarity based on MF. In another study demonstrated in [57], protein-protein interaction databases were used for the assessment of various semantic similarity measures. Five measures were compared: three content-based measures, Resnik's, Lin's, and Jiang and Conrath's, and two graph-based methods - the union-intersection and the longest-shared-path. The union-intersection is the ratio of the number of shared nodes in two induced graphs to the number of all unique nodes, while the longest-shared-path is the length of the longest shared path. However, the union-intersection and the longest-shared-path only consider partial information about the structure of GO graph. Therefore, it is not surprising that Resnik's measure is the best performer when all measures were assessed using human protein-protein interaction data and pathway analysis.

Although Resnik's measure benefits from its simplicity and outperforms others in

many studies [2, 14, 57, 90], none of the semantic similarity measures stands out as having a clear advantage. Besides, the results suggest that the three different aspects of GO are only weakly correlated [2, 90].

Existing GO validity indices [2, 90] transform term-term similarities into gene-gene similarities and furthermore into cluster-cluster similarities. The two-stage transformation unavoidably results in information loss. In the process, gene-gene similarities are often calculated based on the assumption that the average or maximum value of term-term similarity can be used to represent gene-gene similarity. This is problematic, since one cannot expect the average or maximum value to be representative for the whole population. On the other hand, such GO validity indices are often not robust enough against uneven granularity and noises in GO. Because of the noisy and incomplete aspects of GO, semantic similarity is bound to be noisy, which in turn worsens the quality of gene-gene similarity.

Another big concern about this type of method is that the terms used to represent gene functions cannot be compared on the same level. To solve this problem, Barriot *et al.* [6] proposed a mathematical metric for finding the most pertinent terms in gene clusters to represent their functions. However, whether the pertinent terms from two sets of genes can be compared on the same level remains unclear. Meanwhile, none of the assessments performs any test about the fitness of these semantic similarity measures. As it was noted in [90], the ability of the above validation techniques to rank clustering algorithms in terms of their feasibilities in biological prediction remains debatable.

#### 3.3.3 Experimental Assessment

We assess existing validation methods based on GO semantic similarity by a simple standard: their abilities to differentiate between random and meaningful partitions. The term-term similarity are calculated using three measures, Resnik's, Lin's, Jiang and Conrath's measure. Gene-gene similarity are computed based on average term-term similarity. Three existing cluster validity indices, Silhouette index [111], Davies-Bouldin index [33] and the Dunn index [41], are used to evaluate partitions based on semantic similarity [125].

##### 3.3.3.1 Clustering validation indices

###### Silhouette index

Given a set of genes  $\{g_i | i = 1, 2, \dots, n\}$  and a partition of  $P = \{C_j | j = 1, 2, \dots, K\}$ , Silhouette index is defined as follows. For each gene  $g_i$  of cluster  $C_j$ , a confidence measure, the silhouette width  $s(g_i)$ , is defined as

$$s(g_i) = \frac{\min(dB(g_i)) - dW(g_i)}{\max\{dW(g_i), \min(dB(g_i))\}}, \quad (3.3)$$

where  $dW(g_i)$  is the average distance from  $g_i$  to all other genes in cluster  $C_j$  and  $dB(g_i)$  is the average distance between  $g_i$  and all genes in other clusters  $C_k$ ,  $k \neq j$ . Gene assignments with a large  $s(g_i)$  (almost 1) are very well clustered, a small  $s(g_i)$  (around 0) means that the gene lies between two clusters, and assignments with a negative  $s(g_i)$  are probably placed in the wrong cluster. Thus, the overall quality of a partition  $P$  can be measured using

$$S(P) = \frac{1}{n} \sum_{i=1}^n s(g_i). \quad (3.4)$$

#### Davies-Bouldin index

The Davies-Bouldin index aims to identify sets of clusters that are compact and well separated. It is defined as

$$DB(P) = \sum_{i=1}^K \sum_{j=1, i \neq j}^K \max \left\{ \frac{\Delta(C_i) + \Delta(C_j)}{\delta(C_i, C_j)} \right\}, \quad (3.5)$$

where  $\Delta(C_i)$  and  $\Delta(C_j)$  represent the inner cluster distance of cluster  $C_i$  and  $C_j$  and  $\delta(C_i, C_j)$  denotes the distance between the clusters  $C_i$  and  $C_j$ . Usually  $\Delta(C_i)$  are calculated as the sum of the distances of individual genes to the respective cluster centres, and  $\delta(C_i, C_j)$  as the sum of distances between two cluster centres.

#### Dunn index

The Dunn index is defined as the ratio of the smallest distance between observations not in the same cluster to the largest intra-cluster distance. It aims to maximise inter-cluster distance and minimise intra-cluster distance. This index is to identify clusters that are compact and well separated, defined as

$$D(P) = \min_{1 \leq i \leq K} \left\{ \min_{1 \leq j \leq K; j \neq i} \left\{ \frac{\delta(C_i, C_j)}{\max_{1 \leq l \leq K} \Delta(C_l)} \right\} \right\}, \quad (3.6)$$

with  $\Delta(C_i)$  and  $\Delta(C_j)$  having the same meaning as they have in the Davies-Bouldin index.

#### 3.3.3.2 Experiment

In this experiment, the semantic similarities for term-term similarities are computed separately on three GO ontologies, BP, MF, and CC. Their averages are used as gene-gene similarities and as input (distance measurements) to the three cluster validity



indices, Silhouette index, Davies-Bouldin index and Dunn index.

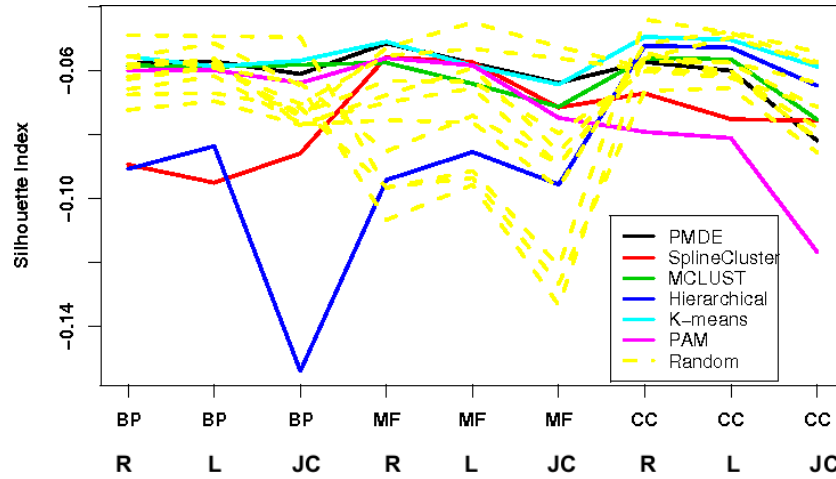


Figure 3.3: Experiments on discriminating random partitions (yellow curves) from meaningful partitions (non-yellow curves) with semantic similarity based on the Silhouette index. For each of the GO category, three semantic similarity measures, Resnik's (R), Lin's (L), Jiang and Conrath's (JC) measure, are used.

For Davies-Bouldin index and Dunn index, there is a choice of linkage methods when computing the inter-cluster distances and intra-cluster distances. For inter-cluster distance there are choices of complete and average linkage, and for intra-cluster distance there are choices of complete, average and Hausdorff linkage [66]. In total, there are six linkage combinations for the computation of both Davies-Bouldin index and Dunn index. First, six clustering methods, including PMDE (as proposed in Section 2.3), SplineCluster (as described in Section 2.2.3), MCLUST (as described in Section 2.2.4), Hierarchical cluster, K-means clustering and Partitioning Around Medoids (PAM), are applied to the yeast Y5 data set as described in Section 2.4.2. Then the three validity indices, Silhouette index, Davies-Bouldin index and Dunn index, are applied on the six resulting partitions and 10 random partitions.

The results from silhouette index are plotted in Figure 3.3. The results based on six linkage methods for Davies-Bouldin index and Dunn index are plotted in Figure

### 3.3 Existing Methods Based on GO Semantic Similarity

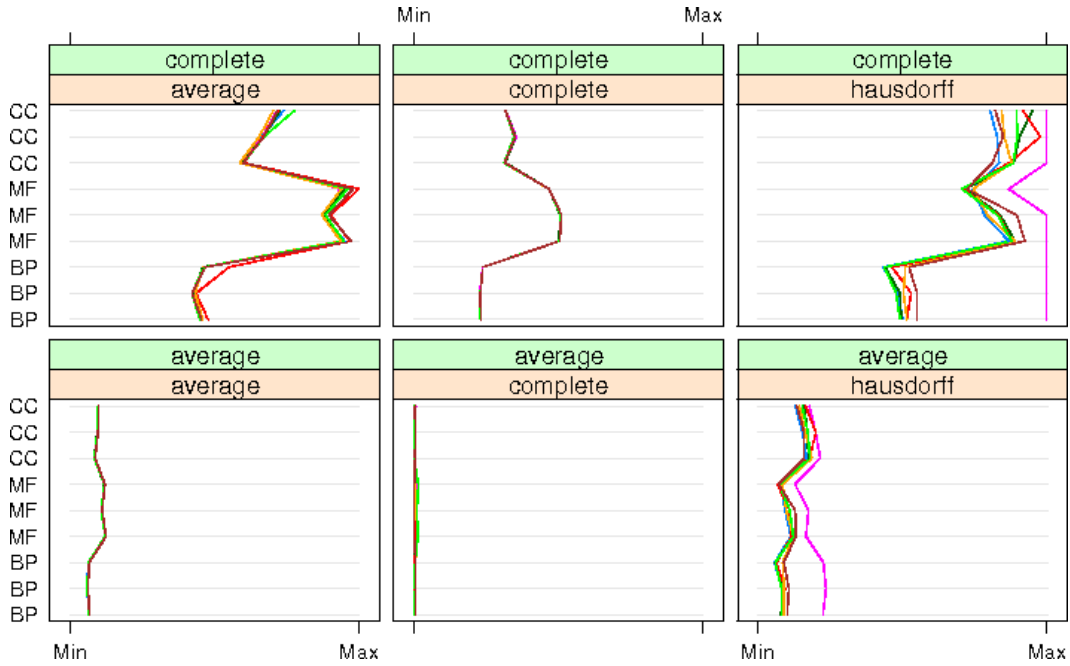


Figure 3.4: Experiments on discriminating random partitions (yellow curves) from meaningful partitions (non-yellow curves) with semantic similarity based on the Davies-Bouldin index. Colour codes are provided in the legend in Figure 3.3.

3.4 and 3.5, respectively. While the green box corresponds to the linkage method for inter-cluster distance computation and the orange box for intra-cluster distance computation. Curves are colour coded for the identities of clustering methods which remain the same in all experiments, with the legend in Figure 3.3. In essence, the objective of this experiment is to see if the random partitions (yellow curves) can be differentiated from other valid partitions (non-yellow curves) by the validity indices. From Figure 3.3, 3.4 and 3.5, it is clear that although occasionally the indices pick up perhaps exceptional partitions, none of them can differentiate the random partitions, based on all three semantic similarity measures. Hence, the ability of the existing GO-driven validation techniques to rank clustering algorithms based on semantic similarity remains unclear.

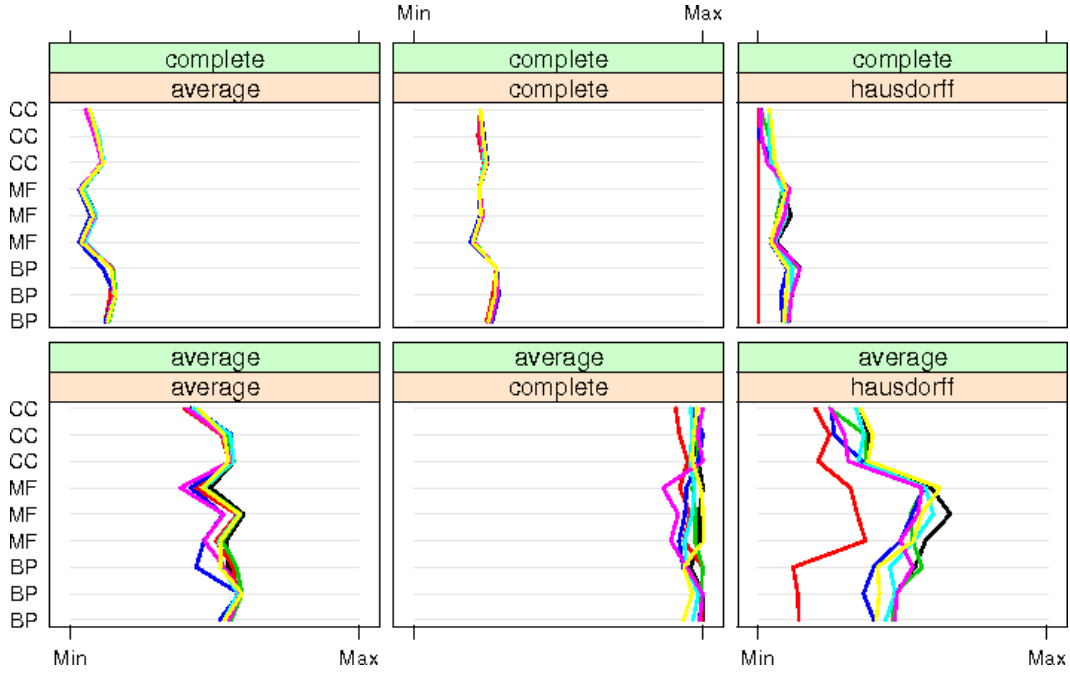


Figure 3.5: Experiments on discriminating random partitions (yellow curves) from meaningful partitions (non-yellow curves) with semantic similarity based on the Dunn index. Colour codes are provided in the legend in Figure 3.3.

### 3.4 Proposed Validation Method

Although the afore-mentioned three GO-driven indices are not able to provide effective solution for clustering validation, GO possesses useful information worth tapping. In this section we introduce a clustering validity index with two sub-indices: within-cluster compactness (WCC) and between-cluster similarity (BCS), upon the establishment of a new distance measure between GO terms. Before that, two important statistics to be used in the proposed index,  $p$ -value and information content, are reviewed.

For each cluster  $C_k, k \in \{1, 2, \dots, K\}$  in a  $K$ -cluster partition, the hypergeometric distribution [131] can be used to identify over-represented GO terms  $T_k = \{t_i | i = 1, 2, \dots, L\}$  in one of the three GO categories, with  $L$  the total number of over-represented terms. Their corresponding  $p$ -values  $\{p_i | i = 1, 2, \dots, L\}$  are calculated as Eq.(2.27). The lower

the  $p$ -value is, the more unlikely is the null hypothesis that a term appears by chance hence the more significant a term is. The set of over-represented GO terms  $T_k$  are of interest since they represent the most common functions shared by genes within cluster  $C_k$ . Following, an induced GO relationship graph  $G_k$  for cluster  $C_k$  can be constructed using  $T_k$  as leaves, linking to all their ancestors until one of the three root ontology terms (BP, MF, and CC) is reached. Since an induced GO graph can be obtained using a certain number of over-represented GO terms from each cluster,  $K$  clusters can then be mapped to  $K$  induced GO graphs. GO graphs thus provide straightforward representation of the functional groups within a set of genes.

Another important notion is information content (IC). Let  $IC(\cdot)$  denotes the information content of a term. While  $p$ -value measures the biological relevance of a term to a specific gene cluster, information content can indicate the specificity of a term regarding the whole population. Although it has been pointed out that not all of the less frequent terms are informative [99], this criterion can serve as a general guideline if no prior information is available. Nonetheless, users should use evidence codes of their choices when computing information content. The information content of a term  $t$  is defined as the negative logarithm of the probability of observing the term or its offsprings in one of the GO categories, i.e.,

$$IC(t) = -\ln(freq(t)/freq(root)), \quad (3.7)$$

$$freq(t) = annot(t) + \sum freq(children(t)), \quad (3.8)$$

where  $annot(t)$  is the number of genes annotated with term  $t$ ,  $children(t)$  is the set of all children terms of  $t$ . Therefore, information content has a minimum value of 0

for the root term and a maximum value of  $-\ln(1/O) = \ln(O)$ , where  $O$  is the sum of occurrences of all terms in this GO category.

#### 3.4.1 GO-based Term-Term Distance

Since GO is a directed acyclic graph, uneven granularity and biological relevance of certain terms need to be considered when evaluating the distance between two distinct GO terms. Biological relevance of certain terms for a specific set of genes can be measured using the  $p$ -value while information content indicates biological specificity of a term. To overcome the limitations discussed in Section 3.1, evaluation can take the GO structure, the height of the graph and the number of branches into account. To this aim, graph theory will be useful in constructing a mathematical GO measure.

First of all, to provide a functional distance measure between terms, we propose a graph-based strategy. A well-defined mathematical measure of term-term distance is of crucial importance. It enables predictions of relationships between gene clusters that would have been impossible if the GO structures could only be compared empirically. Of the many paths existing between two terms, the shortest path  $sp(t_i, t_j)$  between two terms,  $t_i$  and  $t_j$ , is defined as the path through which the two terms first reach a shared parent, the lowest common ancestor (lowest common ancestor).  $sp(t_i, t_j)$  is computed with Dijkstra's algorithm [37]. Since GO is a directed acyclic graph, uneven granularity and biological relevance of certain terms should be considered when evaluating the distance between two distinct GO terms. For example, the depth of GO reflects mostly the rank in categorisation rather than the intrinsic properties of terms. Therefore, instead of treating all edges on the same scale, we assign edge weights to all edges along the path. The idea is that the distance from a term to a more specific child term should

be larger than it is to a more general child term. This results in the definition of edge weight between two terms  $t_p, t_c$

$$w_{cp} = 1 - IC(t_p)/IC(t_c), \quad (3.9)$$

where  $t_p$  is the parent of  $t_c$  in a GO graph. Since the information content of a parent term is no higher than that of a child term, edge weights defined in Eq.(3.9) are bounded in  $[0, 1]$ . In the case of a parent term and a child term having the same information content value, the edge weight is 0. When  $t_p$  is a root term, the edge weight is 1. In this sense, the edge weight reflects the difference between a parent term and a child term in the sense of biological specificity. For terms that share the same parent, the more specific a child term is, the higher its information content is, thus the larger the edge weight is.

Given a graph structure of GO as described above, we now define the term-term distance  $d_{ij}$  between  $t_i, t_j$  and  $d_i$  between  $t_i$  and the root term, with  $x, y$  denoting the nodes along the shortest path, as:

$$d_{ij} = \begin{cases} 1 + \sum_{edge(x,y) \in sp(t_i, t_j)} w_{xy}, & t_i \neq t_j \\ 1, & t_i = t_j, \end{cases} \quad (3.10)$$

$$d_i = 1 + \sum_{edge(x,y) \in sp(t_i, root)} w_{xy}, \quad (3.11)$$

with  $sp(t_i, t_j)$  the shortest path between two terms,  $t_i$  and  $t_j$ . The latter case of  $d_{ij}$  is more likely to happen in the situation when the same term is represented in two gene clusters, for which assigning a constant value 1 to this case helps introduce a penalty, as shown later in Section 3.4.3. In summary, this functional distance measure reflects

the relevance details of all terms along the path and the graph structure of the induced GO graph.

### 3.4.2 Within-Cluster Compactness

Intuitively, a functionally compact GO graph for a gene cluster is characterised as a deep and narrow graph without wide spreading subgraphs. Deep graph indicates specificity in over-represented gene functions, while subgraphs represent different functional groups. This can be computationally described as a result of long distances between over-represented GO terms and root term, and short distances between the over-represented terms. For example, in Figure 3.2, the two big subgraphs with terms ‘GO:0009987’ and ‘GO:0008152’ at top represent two main functional groups in this gene cluster. This should result in low score in functional compactness.

We propose Functional Compactness (FC) to describe the level of compactness of the functional cluster as described above, and an index, Within-Cluster Compactness (WCC), to combine FC for all clusters in order to summarise the overall compactness of a partition. A large value of  $FC$  indicates a functionally compact cluster.

Given a  $p$ -value cut-off  $\rho$ , GO terms  $t_i, t_j$  with corresponding  $p$ -values  $p_i, p_j$  lower than  $\rho$  are selected. Meanwhile, the measure should be normalised to the sizes of clusters and indicate the significance of a term regarding to its  $p$ -value. FC for a cluster  $C_k$  is defined as

$$FC_\rho(C_k) = \frac{\sum_{t_i \in T_k} d_i \cdot (\log p_i)^2}{\sum_{t_i \in T_k} \sum_{t_j \in T_k, j \neq i} d_{ij} \cdot \log p_i \cdot \log p_j}. \quad (3.12)$$

Summing up the distances between over-represented terms to the root term, the numerator in FC formula credits deep and narrow graph. The denominator suppresses

cluster with loosely related terms, since the longer the distances between terms are, the less functional compact a cluster is. In other words, FC discourages subgraphs by involving long distances between terms in two subgraphs. Notably, if a cluster is not significantly enriched, e.g., for a certain  $p$ -value cut-off  $\rho$  it has less GO terms that will contribute to FC, such a cluster also scores lower. Combining FC of all clusters  $C_k$  in a partition  $P$ , WCC can be defined as

$$WCC_\rho(P) = \frac{\sum_{k=1}^K \ln |C_k| \cdot FC_\rho(C_k)}{\sum_{k=1}^K \ln |C_k|}, \quad (3.13)$$

where  $\ln |C_k|$  is the natural logarithm of the size of cluster  $C_k$  and  $\sum_{k=1}^K \ln |C_k|$  serves as a normalising factor. The purpose of involving  $\ln |C_k|$  is to remove the effect from the cluster size.  $WCC_\rho$  serves as a measure for a clustering outcome in terms of its compactness in functional representation.

#### 3.4.3 Between-Cluster Similarity

The idea behind the proposed Between-Cluster Similarity is that the overlapping degree between two graphs can indicate their functional similarity. To computationally depict the overlapping degree between two clusters, we define Functional Similarity (FS) as an indication of similarity/disimilarity (overlap/separation) in terms of biological functions. We also define Between-Cluster Similarity (BCS) which combines the FS scores for all clusters in order to indicate the overall separation among clusters. A large value of  $FS$  indicates a higher level of similarity, since the overlap between two sets of GO terms are more significant. This leads to the formulation of FS as follows.

For a partition  $C$ ,  $K$  induced GO graphs  $G = \{G_1, \dots, G_K\}$  are constructed from  $K$  sets of over-represented terms  $T = \{T_1, \dots, T_K\}$  from the clusters  $C = \{C_1, \dots, C_K\}$ . The



FS between every two clusters  $C_x, C_y$  is:

$$FS_{\rho}(C_x, C_y) = \frac{\sum_{t_i \in G_x} d_i(G_x) \cdot (\log p_i)^2 + \sum_{t_j \in G_y} d_j(G_y) \cdot (\log p_j)^2}{\sum_{t_i \in G_x} \sum_{t_j \in G_y} d_{ij}(G_x \cup G_y) \cdot \log p_i \cdot \log p_j}. \quad (3.14)$$

$d_i(G_x)$  has the same physical meaning as  $d_i$  in Eq.(3.10), with  $G_x$  in the brackets indicating the GO graph identity. The numerator of Eq.(3.14) represents the sum of the sizes of two graphs by summing up the distances between the terms and the root term. The denominator describes the overlap between two functional clusters with the sum of distances between terms in the joint graph ( $G_x \cup G_y$ ). The bigger the overlap in the two functional clusters is, the smaller the distances between terms are in the denominator, hence the higher value of FS. As a summary, for the overall partition, BCS can identify functionally well separated clusters with the definition:

$$BCS_{\rho}(P) = \frac{\sum_{x \neq y} \ln |C_x| \cdot \ln |C_y| \cdot FS_{\rho}(C_x, C_y)}{\sum_{x \neq y} \ln |C_x| \cdot \ln |C_y|}. \quad (3.15)$$

As the name indicates, the smaller this index is, the clusters share less commonality in gene functions, and therefore the better the corresponding partition is.

#### 3.4.4 Combined Index WB

Based on the user-selected GO category/categories, a clustering algorithm's validity measure WB can be calculated by pooling different  $p$ -value cut-offs  $\rho$ . For more than one GO category, the formula of WB takes an additive form so selected GO categories can be linearly combined. For example, if all three GO categories (BP, MF and CC)

are chosen, WB is calculated according to the following formula:

$$WB(P) = \frac{\sum_{\forall \rho} (WCC_{\rho, MF}(P)^2 + WCC_{\rho, BP}(P)^2 + WCC_{\rho, CC}(P)^2)}{\sum_{\forall \rho} (BCS_{\rho, MF}(P)^2 + BCS_{\rho, BP}(P)^2 + BCS_{\rho, CC}(P)^2)}. \quad (3.16)$$

The reason of using a square form is to stress any strong relationship in the GO categories. WB provides a single quantitative measure to facilitate easy comparison of different partitions. The larger WB measure is, the better a partition is since the clusters are compact and well separated.

#### 3.4.5 Confidence Thresholds

In order to draw a statistical conclusion, it is crucial to select representative  $p$ -value cut-offs so that performance can be evaluated on a significance basis. Adjustment of  $p$ -values for multiplicity is performed using the notion of false discovery rate (FDR) [9]. FDR suggests a different point of view when considering testing errors, by controlling the expected proportion of erroneous rejection of the null hypotheses

$E[|False Positives|/(|False Positives| + |True Positives|)]$ . For a given threshold  $\alpha$ , the Benjamini Hochberg procedure states that if  $p_1, p_2, \dots, p_m$  are the observed  $p$ -values, one can find the largest  $b$  so that  $b = \max\{i | p_i \leq i\alpha/m\}$  and reject null hypotheses  $H_1^0, H_2^0, \dots, H_b^0$ . After adjustment,  $p$ -values can be compared directly with any chosen significance level  $\alpha$ .

## 3.5 Experimental Results

Consistency, accuracy and discriminability are the main attributes of the validity indices to be accessed in this experimental section. To this aim, we design three comparative experiments, allowing the proposed WB index to be assessed in many aspects. Biological data sets with distinct features and various complexities are used. Five other validity indices, including two GO-driven and three data-driven indices, are used to compare with the proposed index. Six popular clustering algorithms are selected to represent the wide spectrum of clustering methods.

The three data sets used in the experiments are: yeast cell cycle (Y5) data set (as described in Section 2.4.2), yeast galactose data set (as described in Section 2.4.3), and *Arabidopsis L. Heyn*th diurnal data set. The yeast Y5 data set is popular in the clustering literature for its easy accessibility. The challenges from this data set are posed partly by the ambiguities among the five cell cycle phases and partly by the poor quality of the data set. Compared with Y5 data set, Yeast galactose data set show more distinguishable expression patterns. Its genes reflect four functional categories in GO.

### ***Arabidopsis L. Heyn*th diurnal data set**

The *Arabidopsis L. Heyn*th diurnal data set [124] is collected from an experiment to investigate the impact of the diurnal cycle of the starch metabolism in the leaves of *Arabidopsis L. Heyn*th. It is a larger data set with 800 genes but with only 11 time points and two replicates. For the assessment of our validation scheme, a subset of 800 genes is used which is previously selected using the periodicity test [147]. All data sets in the experiments are filtered. Because of noise and limited annotation knowledge, involving a whole data set prevents us from interpreting the performance of the

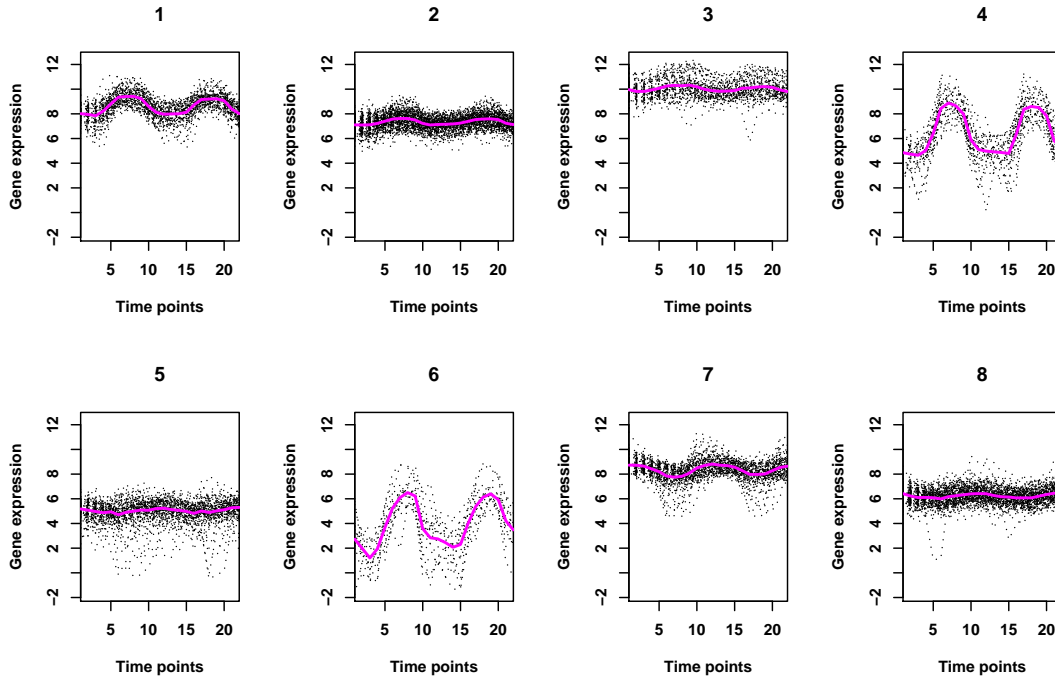


Figure 3.6: The *Arabidopsis L. Heynth* diurnal data clustered into eight clusters by K-means clustering.

proposed methods under evaluation. By using filtered data sets, the interference of unknown factors is significantly reduced, which provides a clearer picture about the role the methods play. Figure 3.6 shows the time series with one replicate concatenated with the other. Ambiguities, especially in the fifth cluster indicates difficulty in this data set in terms of clustering.

In addition to the proposed index, two GO-driven indices are used for comparison: the biological homogeneity index (BHI) and biological stability index (BSI). On the other hand, the three data-driven indices, namely the Calinski and Harabasz (CH) index [21], the Davies-Bouldin index and the Dunn index (as described in Section 3.3.3.1), can be employed to judge the clustering quality from the aspect of data without taking GO into account. The idea behind the CH index is to compute the pairwise sum of

squared distances between clusters using microarray data, and compare that to the internal sum of squared distances for each cluster.

For both the CH index and the Dunn index, a large score corresponds to a good partition. However, for the Davies-Bouldin index, a set of compact clusters is associated with a small value. In the following experiments, the scores of the Davies-Bouldin index are inverted so that large scores correspond to good partitions for all the indices.

We design three experiments to assess the performance of the proposed GO validation indices from different aspects. In the first experiment, six clustering algorithms are evaluated in their applications to the yeast Y5 data set and the Arabidopsis diurnal data set with the six validity indices. In the second experiment, we use yeast galactose data set and its cluster assignment to the four functional categories in a perturbation test to assess the sensitivity and consistency of the proposed validation index with different levels of random errors. The last experiment tests the accuracy of the proposed index by finding the optimum number of clusters for the yeast Y5 data set.

### 3.5.1 Evaluation of Six Clustering Algorithms

We select three model-based and three heuristic clustering methods to be evaluated by the validity indices. PMDE clustering algorithm as introduced in Section 2.3 is a tight clustering algorithm with the capability of detecting outlier/scattered genes. SplineCluster [63] is an efficient hierarchical clustering program based on a spline model with a marginal likelihood criterion. MCLUST [44] is a widely-used model-based method which selects Gaussian models from a pre-defined set and fits them to the data. They are compared with hierarchical clustering (complete linkage), K-means clustering and Partitioning Around Medoids (PAM) [77]. Since both K-means and

PAM are sensitive to initial values, 10 random initialisations are given to both methods and the optimum results are selected by the CH measure.

The ultimate aim of this section is to assess the validity indices. The experiments only show the clustering algorithms' performance in certain cases, with fixed numbers of clusters. Since a clustering algorithm needs to be scrutinized from various angles, the experiments here cannot serve as an overall evaluation of a clustering algorithm. Once a validity index is established as useful, it can then be used to assess clustering algorithm in a more comprehensive setting.

The evaluation of validity indices through the comparative experiment is based on two criteria. First, biological validity index evaluates the ability of a clustering algorithm to produce biologically meaningful clusters. Therefore, a good index should differentiate meaningful partitions from random ones. For each of the data sets, six partitions from the clustering algorithms as well as ten random partitions are generated for comparison. Second, when a GO-driven index agrees with data-driven indices or a majority of indices, it is likely that the judgment for this partition is correct, since it is based on both experimental observations and existing biological knowledge. Hence the corresponding GO-driven index performs accurately. Consequently, good agreement with data-driven indices can serve as positive evidence for GO-driven indices. However, when such connection cannot be found, the partitions may be inspected for their soundness so that validity indices can be assessed.

Since the performance of a clustering method can vary with different data structure and characteristics, experiments are carried out on two data sets of distinct nature, the yeast Y5 data set and the Arabidopsis data set.

### 3.5.1.1 Experiments on yeast Y5 data

The procedure of clustering Y5 data set by various algorithms has been described before (Section 2.4.2.1). However, we use a simpler clustering procedure but with the addition of random partitions, since the focus is on validity index instead of clustering algorithms themselves. For the yeast Y5 data set, five is selected as the number of clusters for all algorithms to represent a simple interpretation of this data set. Six partitions from the clustering algorithms as well as ten random partitions are generated for comparison. Biological validity index evaluates the ability of a clustering algorithm to produce biologically meaningful clusters. Therefore, such an index should differentiate biological meaningful clusters from random ones. We compute the validity scores for six indices for each of the 16 partitions. The three biological indices are based on the GO ontology BP.

Table 3.1: Confidence levels ( $\alpha$ ) and corresponding p-value cut-offs ( $\rho$ ) in the PMDE partition for the Y5 data set.

$\alpha$	0.0025	0.005	0.0075	0.01	0.0125
$\rho$	0.000079	0.000197	0.000356	0.000469	0.000774
$\alpha$	0.015	0.0175	0.02	0.0225	0.025
$\rho$	0.000996	0.001226	0.001639	0.001912	0.002285
$\alpha$	0.0275	0.03	0.0325	0.035	0.0375
$\rho$	0.002598	0.002996	0.003434	0.003738	0.004005
$\alpha$	0.04	0.0425	0.045	0.0475	0.05
$\rho$	0.004371	0.004742	0.005746	0.006238	0.006881

Before the final WB index defined in Eq.(3.16) is compared to other validity index, we first observe the behaviours of individual WCC (Eq.(3.13)) and BCS (Eq.(3.15)) indices to achieve better understanding of the proposed indices. Selected confidence levels and corresponding  $p$ -values cut-offs for the proposed WB index are provided in Table 3.1. Plots of WCC and BCS scores across these cut-offs for each of the

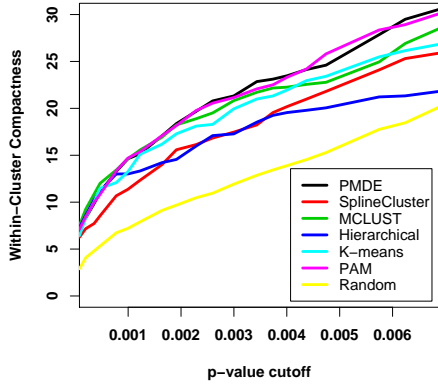
three GO categories are provided in Figure 3.7. From Figure 3.7, we can observe the fairly consistent performance of the proposed indices across different cut-offs  $\rho$ . But incorporating different  $\rho$  into the index is still necessary to provide robust results. Also, users are allowed to define their own selection criterion of the  $p$ -value cut-off  $\rho$  according to their needs, applications, and the organism under study.

Following, validity scores for six partitions and the average score for the 10 random partitions are illustrated in Figure 3.8. On average, scores for PMDE, SplineCluster, MCLUST, Hierarchical, K-means and PAM are 0.93, 0.84, 0.93, 0.86, 0.75, 0.83, respectively. At first glance, PMDE and MCLUST are the best performer for most of the indices, especially in terms of the WB, BSI and Dunn indices. They have the highest average scores. Hierarchical clustering, SplineCluster and PAM also perform reasonably well as judged by most of the indices except the Dunn index. The values from the indices reflect the fact that model-based clustering methods are preferable to heuristic clustering methods such as K-means and hierarchical clustering for this data set. This is reasonable. For the model-based clustering algorithms, PMDE and MCLUST are specifically designed for gene expression time series. Their outstanding performance coincides with established theory [62]. Surprisingly, SplineCluster, also a model-based technique, failed to achieve similar result. Both PMDE and SplineCluster use linear spline model with nonlinear basis functions for data fitting. Nevertheless, PMDE and MCLUST fit one model to each cluster, while SplineCluster fits one model to one time series individually. The approach SplineCluster adopted may lead to overfitting, especially when the time series is short as it is in the case of this data set. On the other hand, PAM demonstrates outstanding quality as a standard technique, although the number of clusters is required as *a priori* knowledge.

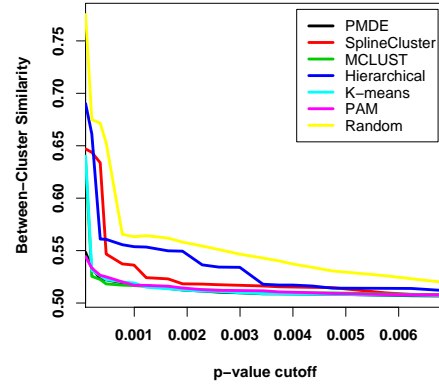
Besides these useful insights about the clustering methods, we also gain better



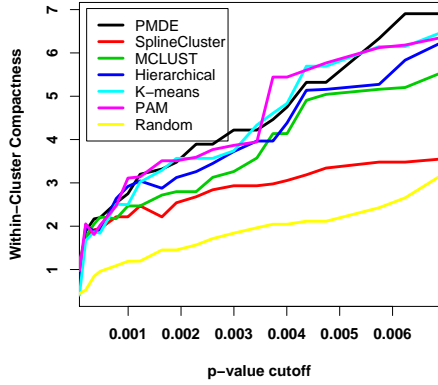
understanding about the validity indices under study. First of all, most indices have the ability to distinguish the random partitions from meaningful partitions. However, BHI scores for K-means and one of the random partitions are almost the same, revealing its deficiency in discriminability. Moreover, BHI scores are often different from other indices. With respect to this index, the best performers are SplineCluster and PAM. On the other hand, the other two GO-driven indices, WB and BSI, are capable of detecting random partitions. They are also more consistent with the data-driven indices, although WB tends to penalise heuristic methods more. At this point, it is still difficult to decide which of WB and BSI outperforms the other.



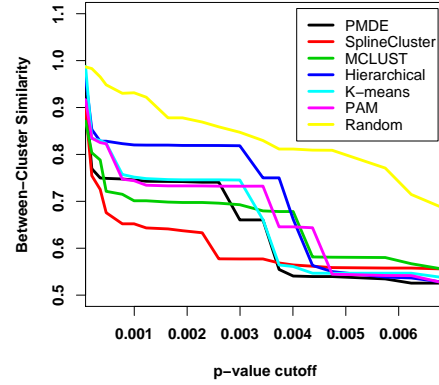
(a)  $WCC_{\rho, BP}$



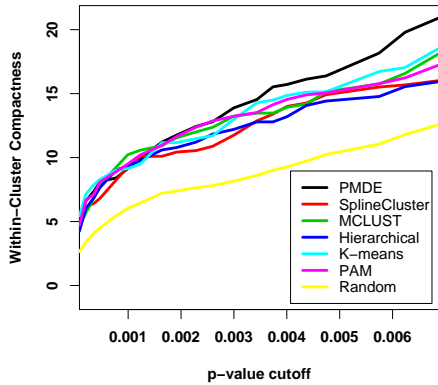
(f)  $BCS_{\rho, BP}$



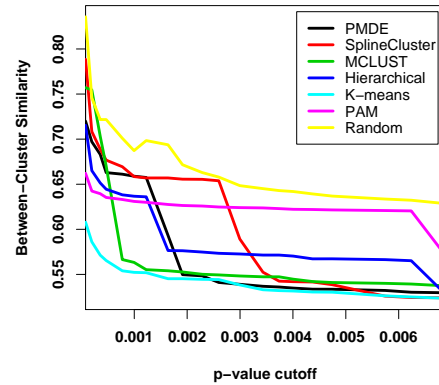
(b)  $WCC_{\rho, MF}$



(e)  $BCS_{\rho, MF}$



(c)  $WCC_{\rho, CC}$



(f)  $BCS_{\rho, CC}$

Figure 3.7: For the Yeast Y5 data set, plots of (a),(b),(c) WCC scores and (d), (e), (f) BCS scores for six clustering algorithms and the average of ten random runs based on the three GO categories BP, MF and CC, respectively.

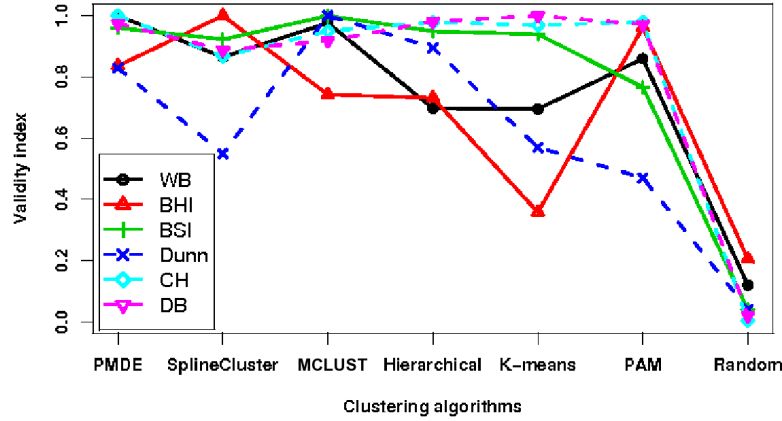


Figure 3.8: For the yeast Y5 data set, normalised scores of six validity indices for various clustering algorithms and random partitions. The solid lines denote that the indices are GO-driven, while the dashed lines denote data-driven indices.

### 3.5.1.2 Experiments on Arabidopsis diurnal data

The Arabidopsis diurnal data set is made up of two experiments. Each experiment consists of measurements at 11 time points of uneven time intervals to capture the periods immediately after the transitions from dark (light) to light (dark). Samples were firstly taken at the end of light period, then at 1, 2, 4, 8, and 12h of darkness and at 1, 2, 4, 8, and 12h of light. For the assessment of our validation scheme, we choose a subset of 800 genes previously selected using the periodicity test [147]. This subset of data was first studied by Rhein and Strimmer for network inference [102].

Consider the sparsity of annotations in Arabidopsis, the short length of time series and the noise in the data, this data set represents a case of higher complexity in our study. Determination of cluster number in this case is more complicated. There is no specific gene selection criterion for choosing the cluster number, unlike the Yeast Y5 data set. Moreover, the number should be selected neither by a validity index nor by a clustering method to avoid bias. While the optimal cluster number selected by

Table 3.2: Confidence levels ( $\alpha$ ) and corresponding p-value cut-offs ( $\rho$ ) in the PMDE partition for the Arabidopsis data set

$\alpha$	0.005	0.01	0.015
$\rho$	8.3e-05	1.0e-03	5.0e-03
$\alpha$	0.02	0.025	0.03
$\rho$	8.3e-03	1.0e-02	1.5e-02

MCLUST is 2, PMDE reports 13 as the optimum. Taking all these into account, we decide on 8 as the cluster number, so that the outcomes of the clustering algorithms is interpretable for the purpose of evaluating of validity indices.

Next, we obtain partitions using the six clustering algorithms. MCLUST often falls in local minimum, yielding singleton clusters. We select the best result with 8 clusters generated from different initialisations. By setting the parameters, PMDE and SplineCluster can also find partitions with 8 clusters. For the biological validity indices, we choose one GO category ‘biological process’ for clustering validation according to the purpose of this microarray experiment. Selected confidence levels and corresponding  $p$ -values cut-offs for the WB index in the PMDE partition are provided in Table 3.2. As can be seen from the table, for a bigger data set such as the Arabidopsis data set, less significant levels  $\alpha$  can be used to reduce the computation cost.

For all validity indices, the scores across the six clustering algorithms are plotted as curves in Figure 3.9. The result appears to be different from the previous experiment for the yeast Y5 data set (c.f. Figure 3.8). As can be seen, hierarchical clustering is judged as the best performer in terms of BHI, DB and Dunn indices, while with respect to WB and BSI K-means clustering is the best. On the other hand, MCLUST receives lowest scores from almost all indices except BHI, which gives its lowest score to SplineCluster. Interestingly, the situation seems to be completely reversed from the previous data set. All indices indicate that better performers are heuristic or ‘simpler’

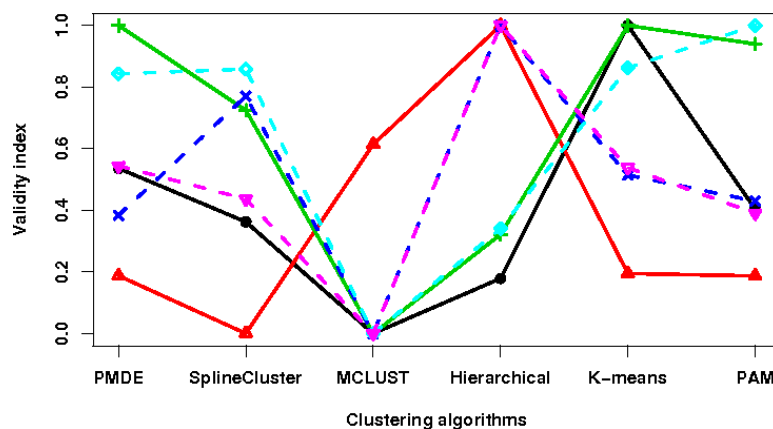


Figure 3.9: For the Arabidopsis diurnal data set, normalised scores of six validity indices for various clustering algorithms. Colour codes indicating the validity index identities are the same as they are in Figure 3.8.

clustering methods instead of model-based methods which dominate the evaluation for the yeast Y5 data set. This is presumably due to the fact that the parameters of model-based methods have not been carefully adjusted. It is generally known that model-based clustering algorithms enjoy full probabilistic modeling and higher level of robustness. However, in some cases they may fail in practice due to the sensitivity to the model assumption or local optimums. The short length of time series in this case and the large number of variables involved makes it particularly challenging for model-based methods. Besides, model-based methods often need special care in implementation to avoid issues such as singularity and local optimum. Their sensitivity to parameter settings such as the cluster number also need to be taken into account.

As far as the validity indices are concerned, there are less connections we can establish between the GO-driven indices and the data-driven indices than there are in the case of Y5 data set. Intuitively, good agreement with data-driven indices can serve as evidence supporting GO-driven indices, if the data set is well-annotated. However, for the Arabidopsis data set this may not be the case, since annotations are far

more sparse and less reliable. Notably, contradiction between GO-driven validation and data-driven validation may partly originate from the noisy nature of GO. However, noise in microarray data itself is another source of errors for data-driven indices. Therefore, consistent high scores of GO-driven indices for a clustering algorithm may suggest its superior ability in handling noise in the data.

For the GO-driven indices, scores of WB and BSI have more in common while BHI scores are again very different. However, the best performers judged by BHI are K-means and PMDE, while the scores of WB indicate that only K-means is the best performer. Hence, we inspect the resulting partitions by K-means and PMDE for their biological meanings. Over-represented terms in the K-means clusters and in the PMDE clusters, together with their information content and  $p$ -values, are extracted and listed in Table 3.3 and 3.4, respectively. From the enriched clusters in the K-means partition, specific GO terms of related biological process (starch metabolism) are found. For instance, the clusters are enriched with photosynthesis (with  $p$ -value  $4.9\text{E-}6$ ), circadian rhythm ( $7.7\text{E-}4$ ), starch metabolic process ( $1.1\text{E-}5$ ), isoprenoid biosynthetic process ( $1.5\text{E-}4$ ).

In contrast, for the PMDE partition, the over-represented terms are less specific and the corresponding  $p$ -values are higher, indicating lower significance. For the over-represented terms in the PMDE cluster, average information content is 6.9 and average  $p$ -value is  $9\text{E-}3$ , while for the K-means partition, average information content is 7.1 and average  $p$ -value is  $6\text{E-}3$ . Successfully, the proposed WB index captures this difference, since it takes into account the specificity of GO terms. Overall, this investigation not only reveals useful insights into the data set and the clustering algorithms, but also provides evidence of the superior performance of the proposed WB index.

Table 3.3: Over-represented GO terms in the K-means partition for the Arabidopsis data set

Cluster	GO ID	GO term	<i>p</i> -values	Gene counts	IC
1	GO:0008610	lipid biosynthetic process	3.94E-05	10	6.413522
1	GO:0044255	cellular lipid metabolic process	6.62E-05	12	6.10293
1	GO:0008299	isoprenoid biosynthetic process	0.000155	5	7.717184
2	GO:0009755	hormone-mediated signaling	0.00288	4	7.089956
2	GO:0009605	response to external stimulus	0.007028	9	6.730338
2	GO:0043687	post-translational protein modification	0.010556	13	5.045178
3	GO:0015979	photosynthesis	4.95E-06	9	7.597105
3	GO:0019684	photosynthesis, light reaction	0.000103	6	7.910601
3	GO:0009414	response to water deprivation	0.000164	7	7.689971
4	GO:0048511	rhythmic process	0.000775	3	9.29272
4	GO:0007623	circadian rhythm	0.000775	3	9.29272
4	GO:0009909	regulation of flower development	0.001956	2	8.337209
5	GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	9.75E-05	22	4.171339
5	GO:0016070	RNA metabolic process	0.004393	9	4.806334
5	GO:0006350	transcription	0.007809	11	4.557692
6	GO:0000904	cellular morphogenesis during differentiation	0.02029	1	8.723626
6	GO:0010090	trichome morphogenesis	0.02029	1	8.781895
6	GO:0010091	trichome branching	0.02029	1	9.602875
7	GO:0005982	starch metabolic process	1.11E-05	7	9.19741
7	GO:0044264	cellular polysaccharide metabolic process	0.000128	7	7.492662
7	GO:0005976	polysaccharide metabolic process	0.000128	7	7.465486
8	GO:00061391	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	0.000491	25	4.171339
8	GO:00062591	DNA metabolic process	0.001259	11	6.144267
8	GO:0016458	gene silencing	0.005689	3	7.713742

Table 3.4: Over-represented GO terms in the PMDE partition for the Arabidopsis data set

Cluster	GO ID	GO term	<i>p</i> -values	Gene counts	IC
1	GO:0048511	rhythmic process	0.000775	3	9.29272
1	GO:0007623	circadian rhythm	0.000775	3	9.29272
1	GO:0009909	regulation of flower development	0.001956	2	8.337209
2	GO:0009605	response to external stimulus	0.001549	10	6.730338
2	GO:0009755	hormone-mediated signaling	0.00288	4	7.089956
2	GO:0051641	cellular localisation	0.044049	10	5.905509
3	GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	0.000117	22	4.171339
3	GO:0016070	RNA metabolic process	0.004739	9	4.806334
3	GO:0006350	transcription	0.008494	11	4.557692
4	GO:00061391	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	0.000955	25	4.171339
4	GO:00062591	DNA metabolic process	0.001787	11	6.144267
4	GO:0016458	gene silencing	0.006411	3	7.713742
5	GO:0015979	photosynthesis	3.48E-06	9	7.597105
5	GO:0019684	photosynthesis, light reaction	8.12E-05	6	7.910601
5	GO:0006950	response to stress	0.000711	16	5.344044
6	GO:0000904	cellular morphogenesis during differentiation	0.02029	1	8.723626
6	GO:0010090	trichome morphogenesis	0.02029	1	8.781895
6	GO:0010091	trichome branching	0.02029	1	9.602875
7	GO:0008610	lipid biosynthetic process	0.000117	10	6.413522
7	GO:0044255	cellular lipid metabolic process	0.000228	12	6.10293
7	GO:0008299	isoprenoid biosynthetic process	0.000281	5	7.717184
8	GO:0005982	starch metabolic process	4.36E-06	7	9.19741
8	GO:0044264	cellular polysaccharide metabolic process	5.19E-05	7	7.492662
8	GO:0005976	polysaccharide metabolic process	5.19E-05	7	7.465486



### 3.5.2 Perturbation Experiment

Table 3.5: Confidence levels ( $\alpha$ ) and corresponding p-value cut-offs ( $\rho$ ) in the starting partition for the Galatose data set for the perturbation experiment

$\alpha$	0.0025	0.005	0.0075	0.01	0.0125	0.015
$\rho$	8.14E-28	1.17E-25	1.34E-24	7.22E-23	2.01E-21	0. 6.53E-20
$\alpha$	0.0175	0.02	0.0225	0.025	0.0275	0.03
$\rho$	2.51E-17	1.00E-15	2.67E-14	1.50E-12	0.002598	0.002996

In this experiment, we assess the indices’ consistency over increasing level of perturbation and their sensitivity to small perturbation. By perturbation we mean small error to be introduced into the system currently under evaluation. The yeast galactose data set is selected, both because it is relatively well annotated and that the ground truth, its assignment to four functional categories, is given. Starting with the four true/functional clusters, each time 2 more genes are assigned wrong cluster memberships. Resulting values for the GO-driven indices WB, BHI and BSI are plotted across the perturbations in Figure 3.10(a), while values for WB index and the data-driven measures Dunn, CH, and DB index are plotted in Figure 3.10(b) for clarity. All validity scores are normalised in this chapter to facilitate comparison. The further to the right of the “Perturbation” axis in Figure 3.10, the greater the perturbation level, hence the worse the quality of the partition. So it is expected that a good validity index should associate lower values to partitions corresponding to higher perturbation levels.

The steady decent of WB index is a strong indication of its consistency. It is also consistent with the data-driven indices, which again proves that the partition quality is worsening. We observe that the Dunn index is very sensitive to perturbations. This is reasonable, since the Dunn index uses only the minimum intra-cluster distance and maximum inter-cluster distance, while the CH and the DB take all distances into account. In contrast to the descents of most indices, BHI values tend to increase after

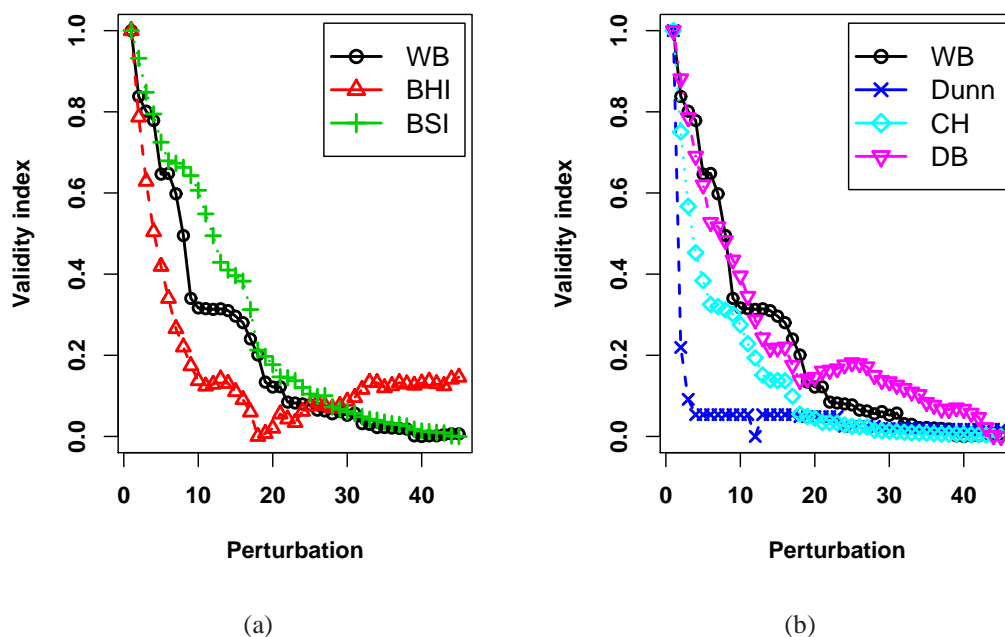


Figure 3.10: Normalised scores of validity indices with increasing level of perturbation in the yeast galatose data set. Large values correspond to good partitions for all the indices. (a) GO-driven validity indices WB, BHI and BSI calculated based on the GO category ‘Biological process’, (b) WB index and data-driven indices.

the 25th experiment. This may be due to the fact that the penalty for such perturbation imposed on BHI is not heavy enough. Another possible reason lies in BHI’s low specificity of GO categories as analysed in Section 3.2. Specific and general GO categories are treated on the same level, but a general functional category may not differentiate true clusters from wrong clusters. For example, a term ‘metabolic process’ covers 191 genes in this data set, thus has no discriminative power. Overall, among the GO-driven indices, better performers in this experiment are BSI and WB.

### 3.5.3 Finding Optimum Number of Clusters

To test the accuracy of the validity indices, we apply them to the yeast Y5 data set to find the optimum number of clusters. While the complexity of this gene expression data set poses acute challenge to the clustering algorithms, the degree of annotation to this data set provides an excellent and accurate basis for the evaluation of biological validity measures. The experiment proceeds as follows. First, partitions with a range of cluster numbers [3 – 12] are obtained for each of the six clustering algorithms. Then the validity scores are computed using all validity indices. We examine the results by each clustering algorithm. Interestingly, only partitions from SplineCluster and hierarchical clustering can provide discriminative evidences for evaluating the advantages/disadvantages of the indices. Although the two algorithms are not the best for this data set from the previous experiment, they provide fairly consistent results across different cluster numbers, while others appear to be sensitive to cluster number. Hence, results based on SplineCluster partitions and hierarchical clustering partitions are depicted in Figure 3.11(a) and (b), respectively.

First of all, all figures (including the ones not provided here) show that CH increases and BSI decreases monotonically, which suggests CH and BSI's sensitivity to cluster numbers. Hence, they fail to achieve the purpose of the test. Although it seems from Figure 3.11(a) and (c) that WB tends to give higher score for smaller cluster numbers, its score for the five-cluster partition stands out. Consider that genes in this data set were originally selected depending on whether their expression peak in one of the five cell-cycle phases, this five-cluster partition may correctly separate the cell-cycle genes. This is further confirmed by the BHI which also selects five as the optimum number. In the same figure, Dunn and DB indices only monotonically go up

and down, respectively. However, these two indices have different performance in Figure 3.11(b) for the hierarchical clustering partitions. In this figure, WB, BHI, Dunn and DB have a preference for the numbers ranging from four to seven. In particular, they get high scores for cluster number five and six. The highest scores for WB and BHI occur when the data set is partitioned into six clusters. Our previous analysis of this data set suggests that it is also possible that this data set has six functional categories (see Table 2.4). Since biological pathways have a hierarchical structure, expressions of genes involved in sub-pathways can be clustered into sub-clusters. Therefore, the hierarchical algorithm may give a good solution when partitioning the data set into 6 clusters. Overall, the only indices that do not have monotonic behaviors across cluster numbers are the BHI and WB, reflecting their potential in selecting optimal number of clusters. As a summary, WB is the only GO-driven index that has excellent performance in all three experiments.

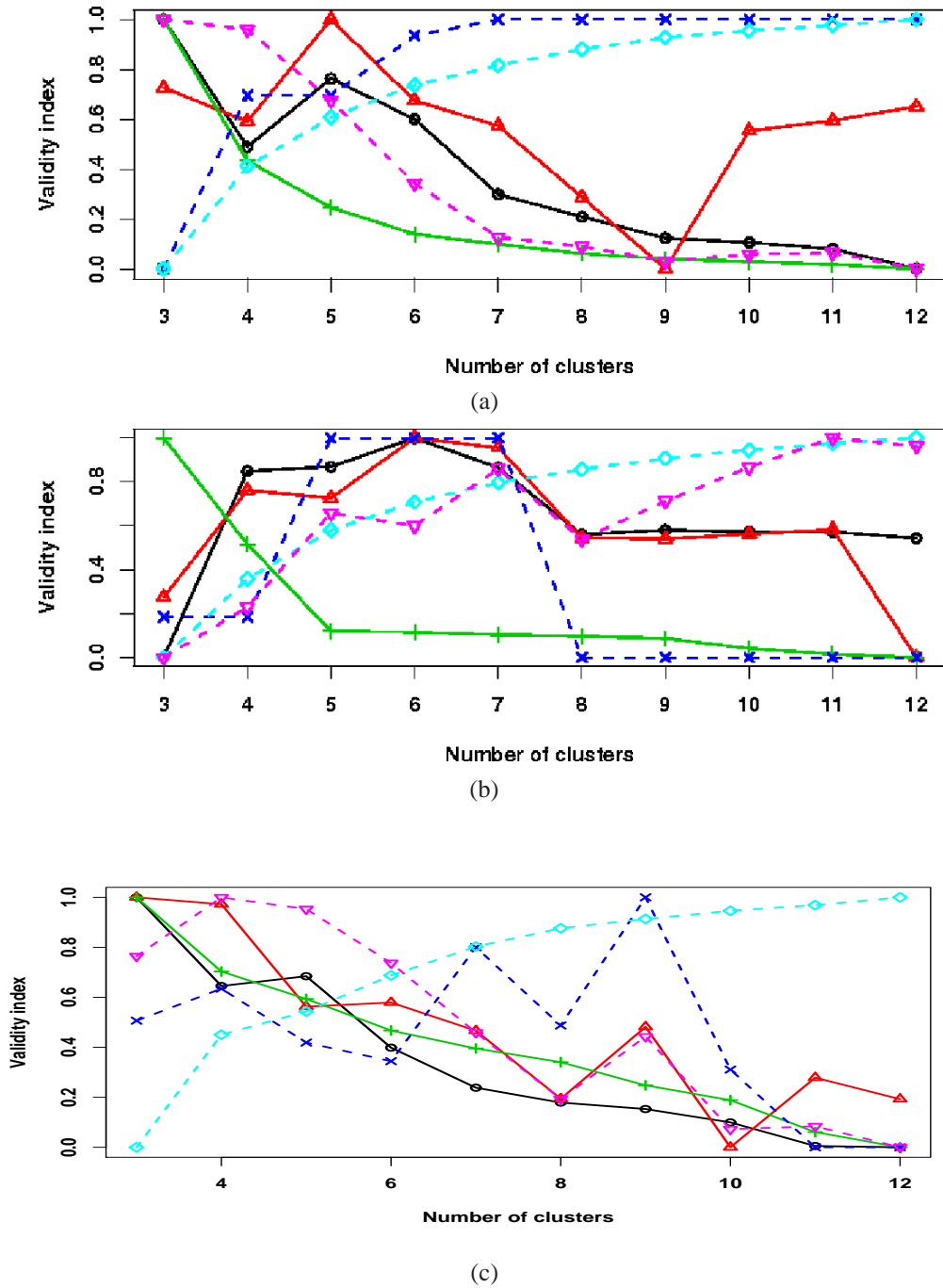


Figure 3.11: Scores of the six validity indices as a function of cluster numbers for the yeast Y5 data set using (a) SplineCluster algorithm, (b) hierarchical algorithm, (c) PMDE algorithm. Colour codes are the same as they are in Figure 3.8 (black: WB, red: BHI, green: BSI, dark blue: Dunn, light blue: CH, pink: DB).

## 3.6 Conclusions

In this chapter, we design a clustering validity index WB to overcome the challenges presented by the complex structure of GO. Based on a new term-term distance defined within the realm of graph theory, the WB index successfully incorporates the strength of relationships between terms. Another desirable feature of the proposed index is that it relieves the assumption that GO terms are compared on the same level. It takes into account not only the variations in biological specificities in GO terms, but also the significance of terms to the gene clusters. Therefore, it is essentially different from the validity measures where GO terms are used as functional categories, such as BHI and BSI. Benefited from these features, the proposed WB index has proven its superior performance in the experimental evaluation.

In the comparative experiments, the proposed WB index's preference for clustering methods provides useful insights into these methods as well as the data sets, and the result coincides with established theory. It also demonstrates its consistency and sensitivity over different levels of random errors. Finally, it proves to be useful for selecting the optimal cluster numbers using biological knowledge. Although BHI and BSI are excellent in some of these aspects, neither of them outperforms WB overall. In summary, this study elucidates much insight into the validity indices, the clustering methods and the data sets. We believe that the proposed index can aid in a more efficient and effective utilisation of the valuable GO information.

With the proposed index, one can select a clustering algorithm that helps reduce data dimension and select key components. However, for the next step of biological inference, for example, discovery of transcriptional regulatory relationships, single source of data is often not sufficient. When more than one biological data source is

available, integrative analysis is likely to be significantly advantageous, and is currently the subject of ongoing research. In the next chapter, other data sources will be combined with gene expression data to increase the confidence in the inference of gene networks, with the help of a new Bayesian method.

## **Chapter 4**

# **A Bayes Random Fields Approach for Integrative Large-Scale Regulatory Network Reconstruction**

### **4.1 Introduction**

One major aim in functional genomics is the reverse engineering of transcriptional regulatory network, which brings the understanding of functional mechanisms in organisms to a higher level. In the hope of discovering transcriptional regulatory activities, one promising research direction is the integrative analysis of diverse data sources [75].

In a transcriptional system, genes and proteins interact with each other in various ways as shown in Section 1.1 and 1.2.3. Basic interactions include transcription factors binding to their target sites on DNA. More complex interactions exist to account for a proportion of all interactions. For example, proteins can combine to form multi-



protein complexes that can perform higher level functions in regulation, for example cleaving RNA or unzipping DNA. These interactions either impose constraints on, or provide capabilities to, the regulatory functions being performed. The interactions involved in transcriptional regulations can be collectively represented as a transcriptional regulatory network.

It is often difficult to conclude which genes play a regulatory role and how genes regulate each other with traditional biology experiments. It is then a major challenge in functional genomics to map out the topological and dynamical properties of the regulatory network. The availability of diverse data from high-throughput experiments has motivated many computational methods, see [25, 75, 135, 148] for good reviews. Naturally, integrating information from different processes and interactions involved in regulatory activities contributes to a deeper understanding of the underlying system, and therefore constitutes a promising direction for regulatory network reconstruction. However, huge amount of data incurs many difficulties in information exploitation, which entails objective techniques.

A key challenge in data integration is the development of a robust system that can be routinely applied to heterogeneous and noisy data. However, such system has not yet been proposed. The reasons are manifold. First, biological data are of quite different nature and formats. For example, microarray gene expression data are often high-dimensional if they are sampled over time, whilst most of the other data types, e.g. protein-protein interaction data, are one-dimensional. The problem becomes how to facilitate effective integration between data of different formats. Another reason is that the coverage of each data type is different from each other. While gene expression data cover almost the entire genome, other data sources are far more sparse. For instance, transcription factor binding data can only partially cover the interactions be-

tween transcription factors and other genes. Therefore, the integrative system has to be carefully designed to address these issues.

In this chapter, we propose a Bayes-Random Fields approach (BRFs) for the integrative analysis of large-scale regulatory networks. The proposed system is capable of integrating unlimited data sources for discovering relevant network architecture of large-scale networks. A potential function is designed to impose a modular constraint on the resultant network, teamed with a full Bayesian approach for combining information from heterogenous data sets. The probabilistic nature of our framework facilitates robust analysis in order to minimise the influence of noise inherent in the data on the inferred structure in a seamless and coherent manner.

Our inspiration comes from the synergy between the problem of regulatory network reverse engineering and the inverse problem in signal processing [130]. Over the past decades, robust statistical methods have matured into some of the most powerful techniques to extract meaningful conclusions from a diversity of data types. The context is similar to the newly arisen study of biological data integration. However, instead of rigidly relying on existing techniques, we aim to take into account the nature of biological data.

This chapter is organised as follows. In Section 4.2, we briefly review the data sources introduced in Section 1.1.1 and discuss rationales and limits in integrative analysis with respect to the features of these data sources. Existing methods for regulatory network reconstruction are reviewed by their categories, bringing up new challenges. The proposed method is then introduced in Section 4.3 and evaluated in experiments with both simulated data sets and *Saccharomyces Cerevisiae* data sets. Further, we provide experimental results and analytical discussions to reveal the varied characteristics of different data sources. It is our hope that such analysis reveals the elementary

structure of regulatory interactions responsible for higher level properties of organisms such as cell growth and death.

## 4.2 Data Sources and Existing Methods

### 4.2.1 Heterogenous Data Sources

The growing availability of genomic, transcriptomic and proteomic data is providing large-scale view of biological systems. With heterogenous data sources available, it is non-trivial to understand the features, relationships and reliability of these data sources for the purpose of regulatory network reconstruction.

The data sources introduced in Section 1.1.1, which are acquired at different stages of cellular activities, relate to each other in one way or another. For example, changes in gene expression may be a direct result of transcription factor binding. In this sense, we can expect information from these data to be combined and form a more powerful prediction system. Indeed, there are many advantages in integration analysis for these data. First, data integration can help filter out erroneous information and increase the confidence in prediction, since biological data are often noisy with many false positives. If there are evidences from multiple independent experiments, reliability of conclusions drawn is greatly improved [75]. Second, data integration can increase the coverage of the genome [135]. Since different data sources may cover different subsets of cellular components, an increase in the coverage in the inference result can be expected by summarising findings from various subsets. Third, integration can help address the problem of specificity in some data sources. For example, gene expression data alone often lack the degree of specificity needed to make accurate

biological conclusions, which can be made up by the transcription factor binding data. The sparsity of transcription factor binding data, on the other hand, can be compensated by the wide coverage of gene expression data. Together, these data sources can help increase overall predictive power from different aspects and on different levels.

There has been, however, a concern about data dependence in the integration literature, which is that subtle correlations and dependencies among data can confound the power of prediction [91]. Recently, Lu *et al.* [91] shed light on this particular aspect by correlating diverse genomic features and observing their integration results. They found no strong dependence in the 16 genomic features studied including gene expression data and functional annotations such as Gene Ontology. Also, it appears that a saturation effect exists in integration systems. At some point, the utility of adding more data sets saturates in the sense that adding more data sets only introduces confusions instead of further reducing noise. By integrating only a few “good” features, maximum predictive power of a system can be achieved. Therefore, the genomic features to be integrated has to be carefully selected. Therefore, it is important to investigate the effects different data types have with respects to the transcriptional regulatory systems under study. In the next subsection, we review existing methods and their choices of data for network reconstruction. Later in the experiments we empirically test the prediction power of the data types studied in this thesis.

### 4.2.2 Existing Methods for Network Reconstruction

In recent years, many researchers devote their work to studying the properties of different genome-scale data, resulting in many methods for reconstructing transcriptional regulatory networks. To understand the essential differences among these methods, it

is important to review existing methods based on the data source/sources they used in order to identify their merits and deficiencies, so that improvement can be made systematically.

### 4.2.2.1 Methods for single data source

Microarray data are perhaps one of the most widely used data sources in this area of research. Many efforts for the reconstruction of transcriptional networks are spent on analysing microarray gene expression data alone [49, 113]. Among earliest works, [50] is an influential paper based on Bayesian networks for gene network inference from gene expression data, with more recent perspectives in [49]. More Bayesian approaches to inferring sparse graphical (Gaussian) models [54] were described in [38, 74].

In more recent years, two types of methods, dynamic Bayesian networks [11] and graphical Gaussian models, account for a major part of research. dynamic Bayesian networks have been widely used in time-series data analysis to account for system dynamics [11, 65, 154]. For example, a dynamic Bayesian networks approach based on a first-order auto-regressive model were applied to gene network reconstruction in [81]. However, inherent problems in dynamic Bayesian networks make them relatively ineffective for large-scale prediction, i.e., when there are many variables. A concern about the inefficiency of dynamic Bayesian networks inspires a number of variant approaches, e.g. a fast “Bayesian-inspired” algorithm by Opgen-Rhein and Strimmer [102].

Graphical Gaussian models are undirected graphical models well known for discriminating direct and indirect correlation between variables. In essence, partial correlation is used as the mathematical foundation for detecting meaningful interactions.

Partial correlation is indicative of direct interactions between a pair of variables/genes, by eliminating the effects from the rest of variables/genes [113]. Previously, graphical Gaussian models have been applied for the reconstruction of gene networks by selecting significant coefficients of partial correlation. Significant coefficients are indicative of direct interactions between genes and therefore represent existing edges in a network. As a breakthrough to solving the small sample problem in gene expression data, Schäfer and Strimmer [113] proposed an shrinkage estimation method of partial correlation and the use of FDR for selecting significant coefficients of partial correlation.

### 4.2.2.2 Methods for multiple data sources

However, single data source is often not sufficient for accurate network modelling [11]. When more than one biological data source is available, integrative analysis is likely to offer significant advantages, and is currently the subject of ongoing research. By integrating multiple data types, one can expect false positives to be reduced and disparities between different levels of the system to be identified. Further, integration helps explain complex biological interactions on a higher level than using a single data alone [11]. Computational techniques have evolved from the simplest voting model [142] to more sophisticated Naïve Bayesian Networks [82, 120], and progressively developed into substantially more complex and powerful systems nowadays [86, 127].

In the integration context, Bayesian methods offer a range of advantages over conventional statistical techniques that make them particularly appropriate for complex and noisy biological data. The Bayesian statistical paradigm is probabilistic in the sense that observations, parameters and hidden variables are treated together in a consistent manner. Consequently, various Bayesian methods for data integration have been

explored for the reconstruction of regulatory networks [11, 19, 76, 82, 127, 137, 162].

Among earliest attempts, [120] set up two probabilistic models for gene expression and protein-protein interaction data, respectively, that can only be solved when unified. Expression data were modelled with Naïve Bayesian networks to define a joint distribution as a product of probabilities of disjoint classes, while protein-protein interaction data were modelled by a binary Markov random fields to represent connections between neighbouring variables.

Later in [51], gene expression data and protein-DNA binding data were jointly considered to infer transcriptional regulatory networks for many chosen yeast transcription factors. However, different data types were not jointly modelled in a coherent framework, and associated measurement errors were not explicitly considered. More complicated integration system was presented by Liu *et al.* [86], where data were jointly modelled within the context specific Bayesian framework for infinite mixture models. In the experiments, the method was able to produce more functionally coherent transcriptional modules than two alternative algorithms, GRAM [5] and SAMBA [128].

Another type of approach uses one data source as prior knowledge to integrate with another in a Bayesian context. For example, Bernard and Hartemink [11] set up dynamic Bayesian networks for modelling gene expression data, combined with transcription factor binding data as prior knowledge and the edge distribution assumption made in [119]. They improved the method in [61] by suggesting a new prior and using dynamic Bayesian networks instead of Bayesian networks so that the network can include cyclic structure. However, the experiment to validate this method was performed on a set of 25 genes with gene expression data consisting of 69 time points, which is far less genes than usually required for network reconstruction nowadays. Sun *et al.*

[127] treated transcription activity represented by expression as a result of transcription factor binding. If the binding data show evidence of regulatory relationships, then the relative binding intensity will be used in modelling the expression of the target gene.

Yet another common approach is to alternate between two data types during the computation process, especially when the main task is to identify regulatory motifs [19, 76, 162]. The strategy to accomplish this involves, first, clustering gene expression data sets, and then isolating the upstream regions of the clustered genes and analysing them for common cis-regulatory motifs. If the identified motifs correspond to known transcription factor-binding sites, the regulatory network that is responsible for the observed transcription state can be inferred.

### 4.2.3 Existing Problems and Prelude to the Proposed Approach

Very often, integrative systems are constrained to two or three different data sources, e.g., gene expression and transcription factor binding data for [11], gene expression and protein-protein interaction data for [100, 121], and gene expression and sequence data for [19, 76, 162]. It is sometimes preferable that the integrative system can be adapted to new data sources. Another research gap is that usually only a small number of genes can be incorporated into a regulatory network, e.g. [11], due to the inefficiency of the learning techniques. However, it is necessary to put the regulatory relationships in a larger context, both because transcriptional activities are usually multi-stages and operate like chain actions involving a large number of genes, and that gene regulations are typically embedded in a vast network of biochemistry interactions [32].

The proposed method to be described later in Section 4.3 differs from previous ones by using a Bayesian framework that can be routinely applied to different data sources,



while remaining efficient enough to facilitate large-scale analysis from the use of partial correlation. Since the focus of this study is the relevant structure of large-scale networks, we only consider undirected graphs. Previous studies have shown that the nature of a network can be recovered even if it is undirected [70]. Moreover, an undirected graph is conceptually simple in the sense that the problem with feedback loops as in a Bayesian network is out of the question, hence is more widely applicable, especially in integrative study when some of the data may be undirectional (for example the gene ontology categories).

The contributions of this chapter are three-fold. First, we propose a full Bayesian approach to incorporate not only microarray time-series data, but also other heterogeneous data sources into a integrative network. Second, we assess the degree to which prediction power increase with the addition of each data source. The effect of integrating heterogeneous data sources is analysed in a substantially more coherent manner. Third, to achieve better understanding about which data source best benefits the integration system, features of heterogeneous data such as specificity and coverage are discussed.

## 4.3 Proposed Bayes Random Fields (BRFs) Integrative Method

The integrative method aims to combine information from heterogeneous data with diverse formats. In integrative study, microarray time-series data attract special attention, because their dynamic features can directly reveal active components within the cell. While the dynamic nature of the data is important, it also incurs challenges because

of the high dimensionality [25]. To reduce data dimension, partial correlation of gene expression time series, for its efficiency and effectiveness, is used as the inference result and incorporated into the integrative framework. Details about partial correlation computation is given in Appendix B.

Since gene expression data is replaced with their partial correlation inference results, the inputs of the integrative system can be unified into probability matrices. Each entry in the matrices can indicate the probability of interaction between a pair of genes, that is, the probability that an edge exists between them in the network. Let  $X$  denote the edges among  $n$  genes in the network  $X = \{x_l | l = 1, 2, \dots, e\}$ , with  $e$  the total number of edges,  $e = n(n - 1)/2$ . Now the integration problem can be formulated as inferring binary variables  $X$  from  $m$  data sets from various data sources, each represented as a matrix of dimension  $n \times n$ . Let  $p(X)$  denote a probability density over hidden variables/edges  $X$ , now we define a Bayes framework with a random fields model for integrative analysis.

#### 4.3.1 Bayes Framework

The aims of the Bayes framework are to integrate information from  $m$  data matrices  $\{\psi_i | i = 1, 2, \dots, m\}$  and to extract regulatory relationships summarised by a common feature  $X$  in the data. Suppose each data matrix represents a property of  $X$ ,  $\{\phi_i | i = 1, 2, \dots, m\}$ , with Gaussian noise  $\{\varepsilon_i | i = 1, 2, \dots, m\}$ , then we have

$$\psi_i = \phi_i + \varepsilon_i, \quad i = 1, 2, \dots, m. \quad (4.1)$$

Now we can set up a model using  $X$  as the common feature/hidden variables among all the data. The objective is to estimate directly from  $\psi$  not only  $\phi$  but also their common

feature  $X$ . The problem can be formulated as

$$\begin{aligned}
 & p(\phi_1, \dots, \phi_m, X | \psi_1, \dots, \psi_m) \\
 &= p(\phi_1, \dots, \phi_m | X, \psi_1, \dots, \psi_m) p(X | \psi_1, \dots, \psi_m) \\
 &\propto p(\psi_1 | \phi_1, X) \dots p(\psi_m | \phi_m, X) p(\phi_1 | X) \dots p(\phi_m | X) p(\psi_1 | X) \dots p(\psi_m | X) p(X) \\
 &\propto p(X) \prod_{i=1}^m p(\psi_i | \phi_i) p(\phi_i | X) p(\psi_i | X).
 \end{aligned} \tag{4.2}$$

Thus in order to get  $p(\phi_1, \dots, \phi_m, X | \psi_1, \dots, \psi_m)$ , we need to define  $p(\psi_i | \phi_i)$ ,  $p(\phi_i | X)$ ,  $p(\psi_i | X)$  and finally  $p(X)$ . The definitions of the first three probabilities are straightforward. Suppose  $\varepsilon_i$  is Gaussian with the mean equal to 0, according to Eq.(4.1) we have

$$p(\psi_i | \phi_i) = \mathcal{N}(\phi_i, \sigma_{\varepsilon_i}^2) = \frac{1}{(\sqrt{2\pi}\sigma_{\varepsilon_i})^e} \exp\left\{-\frac{(\psi_i - \phi_i)^2}{2\sigma_{\varepsilon_i}^2}\right\}. \tag{4.3}$$

There are two classes for the hidden variables  $X$ , 0 and 1, representing the non-existence and existence of an edge, respectively. We can assume that the probability density function in the two classes can be characterized by  $\mathcal{N}(\mu_{i0}, \sigma_{i0}^2)$  and  $\mathcal{N}(\mu_{i1}, \sigma_{i1}^2)$ ,

$$p(\phi_i | X) = \frac{1}{(\sqrt{2\pi}\sigma_{i0})^{e_0}} \exp\left\{-\frac{(\phi_i - \mu_{i0})^2}{2\sigma_{i0}^2}\right\} \cdot \frac{1}{(\sqrt{2\pi}\sigma_{i1})^{e_1}} \exp\left\{-\frac{(\phi_i - \mu_{i1})^2}{2\sigma_{i1}^2}\right\}, \tag{4.4}$$

where  $e_0$  and  $e_1$  denote the number of edges in class 0 and class 1,  $e_0 + e_1 = e$ . With the same principle, we assign  $p(\psi_i|X)$  by using Eq.(4.1), Eq.(4.4) and  $p(\varepsilon_i)$  to the following

$$\begin{aligned}
 & p(\psi_i|X) \tag{4.5} \\
 &= \mathcal{N}(\mu_{i0}, \sigma_{i0}^2 + \sigma_{\varepsilon_i}^2) \cdot \mathcal{N}(\mu_{i1}, \sigma_{i1}^2 + \sigma_{\varepsilon_i}^2) \\
 &= \frac{1}{\left(\sqrt{2\pi(\sigma_{i0}^2 + \sigma_{\varepsilon_i}^2)}\right)^{e_0}} \exp\left\{-\frac{(\psi_i - \mu_{i0})^2}{2(\sigma_{i0}^2 + \sigma_{\varepsilon_i}^2)}\right\} \cdot \frac{1}{\left(\sqrt{2\pi(\sigma_{i1}^2 + \sigma_{\varepsilon_i}^2)}\right)^{e_1}} \exp\left\{-\frac{(\psi_i - \mu_{i1})^2}{2(\sigma_{i1}^2 + \sigma_{\varepsilon_i}^2)}\right\}. \tag{4.6}
 \end{aligned}$$

$p(X)$  is defined in Section 4.3.2.

#### 4.3.2 Random Fields Model

To estimate  $p(X)$ , a random fields model [80] is desirable to represent a known feature of the gene network. A widely accepted concept in transcriptomics is the co-regulation within a gene cluster (co-expression), which can be interpreted as that if a gene is regulating most of the genes in a cluster, it is likely that links also exist between this gene and other genes in the same cluster. In the context of gene network modelling, we define the following model to represent this feature.

To define our clusters, we first perform cluster assignments for genes. The genes are clustered into  $z$  clusters  $\{C_i|i = 1, 2, \dots, z\}$  using gene expression data, preferably by a tight clustering algorithm [157] proposed in Chapter 2 which is designed for gene expression time-series data. The purpose of applying this method is to obtain relatively small/tight clusters, so that relevant information based on these clusters can be inferred by the random fields model. This clustering method is also unsupervised in a sense that the number of clusters needs not be specified. The potential function in the

### 4.3 Proposed Bayes Random Fields (BRFs) Integrative Method

---

random fields model is defined on the edges, i.e., between pairs of genes. Let  $h_{i,C_j}$  be the number of edges between gene  $i$  and cluster  $j$ ,  $d_i$  is the degrees of connectivity for gene  $i$ . The random fields method is formulated as the sum of potentials on all possible edges:

$$p(X) \propto \exp \left[ \sum_i \sum_{j \neq i} \omega_{ij} (h_{i,C_j} + h_{j,C_i} - (d_i/z)^2 - (d_j/z)^2) \right], \quad (4.7)$$

where  $\omega_{ij} = (|C_i| \cdot |C_j|)^{-1}$  is a normalising factor. The first two terms in the potential function represents the number of edges between gene  $i$  and the cluster which gene  $j$  belongs to, and vice versa, while the last two terms are the expected number of edges connecting gene  $i$  and cluster  $C_j$ , and vice versa. The rationale supporting the potential function is that since co-expression indicates co-regulation, a higher potential should be given to the interaction between a pair of genes, if the existing interactions between their clusters are more than expected. An advantage of introducing such dependency is that it imposes a modular constraint as a known gene network feature and iteratively refines the territory currently under evaluation.,

#### 4.3.3 A Gibbs Sampling Algorithm for BRFs

Let  $\theta$  denote the parameter set  $\{\theta_i | i = 1, 2, \dots, m\}$ ,  $\theta_i = \{\mu_{i0}, \mu_{i1}, \sigma_{i0}, \sigma_{i1}, \sigma_{\varepsilon_i}\}$ . Jointly sampling the whole set  $\{\phi_i, \theta_i, X\}$  from large-scale data  $\psi_i$  is intractable. Since now all the variables of interest can be estimated by conditioning on the others, Gibbs sampling can be used to cycle through these conditional statements. By iteratively conditioning on the interim values of all other variables, Gibbs sampler aims to empirically approx-

imate the desired marginal distribution for each variable. We assign *a posteriori* law

$$p(\phi, \theta, X|\psi) \quad (4.8)$$

$$= p(\phi, X|\theta, \psi) p(\theta|\phi, X, \psi)$$

$$= p(\phi|X, \theta, \psi) p(X|\psi, \theta) p(\theta|\phi, X, \psi) \quad (4.9)$$

$$= \prod_i^m p(\phi_i|X, \theta_i, \psi_i) p(X|\psi_i, \theta_i) p(\theta_i|\phi_i, X, \psi_i).$$

Thus given data  $\{\psi_i|i = 1, 2, \dots, m\}$ , the Gibbs sampling algorithm is formulated as the following:

#### 1. Initialisation

- (a) First a random initial value  $X^{(0)}$  is assigned.
- (b) The conjugate priors for the hyperparameter variance  $\sigma_{ik}$  ( $k \in \{0, 1\}$ ) and  $\sigma_{\varepsilon_i}$  in the normal distribution model are the inverse gamma distributions ( $\mathcal{IG}$ ) [53], while for the hyperparameter mean  $\mu_i$  it is given a normal prior. Therefore, first the hyper-hyperparameters  $\alpha_i, \beta_i, \nu_i, s_i^2, \alpha_{\varepsilon_i}, \beta_{\varepsilon_i}, i \in \{1, 2, \dots, m\}$  are assigned. Then the priors are sampled from the following distributions

$$\sigma_{ik}^2 \sim \mathcal{IG}(\alpha_i, \beta_i), \quad (4.10)$$

$$\mu_{ik} \sim \mathcal{N}(\nu_i, s_i^2), \quad (4.11)$$

$$\sigma_{\varepsilon_i}^2 \sim \mathcal{IG}(\alpha_{\varepsilon_i}, \beta_{\varepsilon_i}), \quad (4.12)$$

with  $k \in \{0, 1\}$  representing the two classes of  $X$  values and  $i \in \{1, 2, \dots, m\}$  representing the  $m$  data types.

- (c) Clustering is performed using gene expression data using the unsupervised

tight clustering algorithm proposed in Section 2.3 to obtain  $z$  clusters.

2. For each iteration, sample  $X$  from the posterior distribution:

$$\begin{aligned}\pi(X|\psi, \theta) &\propto p(\psi_1, \dots, \psi_m|X, \theta_1, \dots, \theta_m)p(X) \\ &= p(X) \prod_i^m p(\psi_i|X, \theta_i),\end{aligned}\tag{4.13}$$

which can be achieved according to Eq.(4.6) and Eq.(4.7), respectively. According to Eq.(4.13), for each element  $x_l$  in  $X$ ,  $l = 1, 2, \dots, e$ , two probabilities can be computed:  $p_l^1$  the probability that the element in  $X$  belonging to class 1 and  $p_l^0$  the probability that the element in  $X$  belonging to 0. The probabilities are then normalised and compared with a number drawn from a uniform distribution ( $\mathcal{U}(0, 1)$ ) to decide whether the new value takes 1 or 0. This is to compute

$$x_l = \begin{cases} 1, & \frac{p_l^1}{p_l^1 + p_l^0} \geq t \\ 0, & \frac{p_l^1}{p_l^1 + p_l^0} < t \end{cases}, \quad t \sim \mathcal{U}(0, 1).\tag{4.14}$$

3. Sample  $\{\phi_i|i = 1, 2, \dots, m\}$  from the posterior distribution

The posterior distribution of  $\phi$  is produced by the product of the likelihood function and the prior:

$$\begin{aligned}\pi(\phi_i|\psi_i, X, \theta_i) &\propto p(\psi_i|\phi_i, X, \theta_i)p(\phi_i|X, \theta_i) \\ &= \mathcal{N}(\phi_i, \sigma_{\varepsilon_i}^2) \cdot \prod_{k=0,1} \mathcal{N}(\mu_{ik}, \sigma_{ik}^2) \\ &\propto \prod_{k=0,1} \mathcal{N}\left[\left(\frac{\psi_i}{\sigma_{\varepsilon_i}^2} + \frac{\mu_{ik}}{\sigma_{ik}^2}\right) \cdot \left(\frac{1}{\sigma_{\varepsilon_i}^2} + \frac{1}{\sigma_{ik}^2}\right)^{-1}, \left(\frac{1}{\sigma_{\varepsilon_i}^2} + \frac{1}{\sigma_{ik}^2}\right)^{-1}\right].\end{aligned}\tag{4.15}$$

4. Sample  $\{\theta_i | i = 1, 2, \dots, m\}$  from posterior distributions

$$\sigma_{ik}^2 \sim \mathcal{IG}(\alpha_i + \frac{e_k}{2}, \beta_i + \frac{1}{2} \sum_{X=k} (\psi_i - \mu_{ik})^2), \quad (4.16)$$

$$\mu_{ik} \sim \mathcal{N} \left[ \left( \frac{\nu_i}{s_i^2} + \frac{\sum_{X=k} \psi_i}{\sigma_{ik}^2} \right) \cdot \left( \frac{1}{s_i^2} + \frac{e_k}{\sigma_{ik}^2} \right)^{-1}, \left( \frac{1}{s_i^2} + \frac{e_k}{\sigma_{ik}^2} \right)^{-1} \right], \quad (4.17)$$

$$\sigma_{\varepsilon_i}^2 \sim \mathcal{IG} \left[ \alpha_{\varepsilon_i} + \frac{e}{2}, \beta_{\varepsilon_i} + \frac{1}{2} \sum (\psi_i - \phi_i)^2 \right]. \quad (4.18)$$

5. Repeat Step 2-4 until convergence.

Convergence is determined according to the Zellner and Min criteria [160]. In the case of Gibbs sampling, the unknown parameters can be separated into two sets:  $\{X\}$  and  $\{\theta, \phi\}$ . Therefore we have  $\pi(X, \theta, \phi | \psi) = \pi(X | \theta, \phi, \psi) \pi(\theta, \phi | \psi) = \pi(\theta, \phi | X, \psi) \pi(X | \psi)$ . Let iteration  $b$  be the candidate stopping point of the chain, and  $\pi_b(\hat{x} | \psi)$  be a smoothed empirical estimate,  $\pi_b(\hat{X} | \psi) = \sum_{j=1}^b \pi(X | \theta_j, \phi_j, \psi) / b$ . When the ratio of convergence

$$\hat{\gamma}_b = \frac{\pi(X | \theta, \phi, \psi) \hat{\pi}_b(\theta, \phi | \psi)}{\pi(\theta, \phi | X, \psi) \hat{\pi}_b(X | \psi)} \quad (4.19)$$

is approximately equal to one, we stop the estimation process.

In summary, we empirically obtain the posterior distributions for the parameters and hyperparameters. If the Gibbs sampler has run sufficiently long, this algorithm produces a complete sample of the coefficients. The Gibbs sampler decomposes the whole set of parameters into three sets  $\phi, \theta$  and  $X$ , since the form of random field we have chosen makes an exact sampling of  $p(X | \phi_1, \dots, \phi_m, \psi_1, \dots, \psi_m)$  impossible.

There are a few points we noted here for the proposed algorithm. First, the posterior of mean  $\mu_{ik}$  depends on the data only through the sum of data  $\sum \phi_i$ , meaning that this data summary is sufficient from the data to estimate the unknown mean. Second, as the



data size increases, the value of estimated mean will increasingly depend on the data and variance  $\sigma_{ik}$ , making the prior assumption less important. Last, it is possible that the set genes/proteins with regulatory roles are known for some genomes and therefore regulatory interactions only exist between them and all genes. This can greatly reduce the number of variables and speed up the algorithm,.

## 4.4 Experiments

Both simulated data and biological data are used for experimental evaluation. Biological data can only provide anecdotal evidence in network validation, since the knowledge of gene regulation is far from complete. It seems that we can use functional annotations as golden standard, but annotation information among different annotation databases is too inconsistent to support a large scale evaluation [99]. On the other hand, simulated data sets can provide more controllable conditions to test an algorithm and a standard for benchmarking. However, to obtain meaningful results, the simulated data need to share statistical characteristics with biological data.

For synthetic networks, the proposed algorithm can be compared with graphical Gaussian models on the basis of simulated gene expression data generated by SynTReN. For real gene network, we integrate gene expression data, transcription factor binding data and protein-protein interaction data using the aforementioned framework for yeast *Saccharomyces Cerevisiae*. Comparison of the resulting network and a golden standard network clearly shows the benefits of data integration.

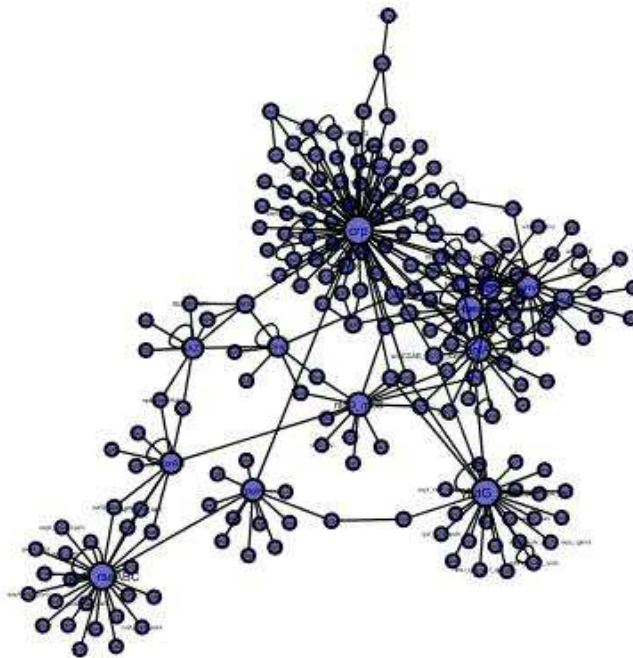
### 4.4.1 Synthetic Networks

We use SynTReN to generate synthetic networks as follows. The topologies of the synthetic networks are sub-sampled from a yeast transcriptional network in [56]. SynTReN uses a sampling method named cluster addition (initial graph is selected as a randomly selected node and all of its neighbors). Combined with external conditions that trigger the network, the expression levels of genes in each experiment are generated according to the activities of their regulators.

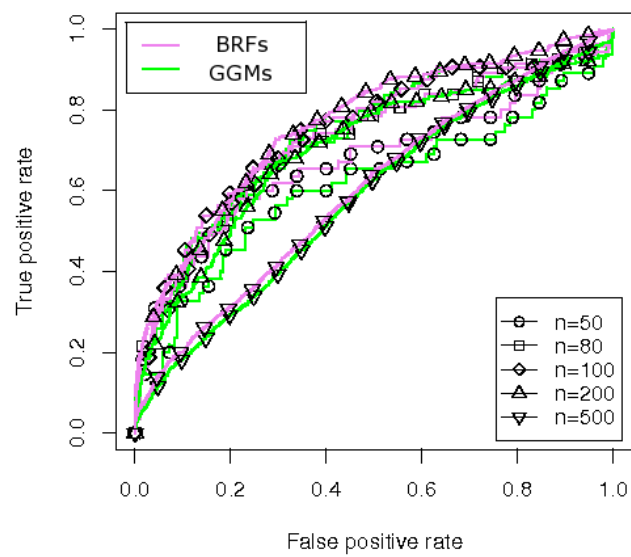
Table 4.1: Some parameter settings for SynTReN software to generate the simulated data sets, the rest are set as default.

Data set	1	2	3	4	5
Background Nodes	50	50	50	50	50
Bio Noise	0.02	0.05	0.08	0.1	0.1
Exp. Noise	0.02	0.05	0.08	0.1	0.1
Noise on correlated input	0.02	0.05	0.08	0.1	0.1
Fraction of complex interactions	0.1	0.1	0.2	0.2	0.3

Five synthetic networks are generated with Gaussian noise (level 15%) and relatively large proportion of complex interaction (30%). Details about the configuration of SynTReN are provided in Table 4.1. The five synthetic networks consist of 50, 80, 100, 200 and 500 genes, respectively. Each network is sampled at 25 time points. A 200-gene synthetic network is plotted in Figure 4.1(a). Note that there is only synthetic gene expression data, so we can compare the results from graphical Gaussian models with those from the proposed BRFs model.



(a) A synthetic network of 200 genes



(b) ROC curves for comparing graphical Gaussian models with the proposed method on five synthetic networks of various sizes.

Figure 4.1: Experimental results for the synthetic networks.

Since with single data source both methods are based on partial correlations, BRFs' performance can be assessed to see the effect of the modular constraint imposed by the random field model. ROC curves for both methods on five data sets are plotted in Figure 4.1 (b). The violet curves representing BRFs inference show superior performance to the green curves representing graphical Gaussian models inference. For the simulated data BRFs make use of its random field component but not the integration feature. In this way, we can observe that the proposed BRFs method improves the results by imposing a modular constraint in network inference.

#### 4.4.2 *Saccharomyces Cerevisiae* Regulatory Network

For the reconstruction of the yeast *Saccharomyces Cerevisiae* regulatory network, three real data sets are integrated in this experiment. The result is compared to a golden standard network to evaluate the accuracy of the proposed method. The three data sets have their unique features: transcription factor binding data provide direct information to understand the regulators involved in transcription; protein-protein interaction data reveal proteins that involved in the same pathway, as well as related to genomic level - interacting proteins are often co-expressed and co-localised to the same sub-cellular compartment. These data types were discussed in detail in Section 1.1.1. Both of the transcription factor binding data and protein-protein interaction are of certain degree of specificity and sparsity, but they can only describe the potential of interaction. In contrast, microarray expression time series are a complementary source that provides dynamic information about the expressions of almost all genes under certain conditions in an organism. Although the data are known to be noisy, they reflect actual interactions in the biological process under analysis.

For gene expression data, the alpha factor arrest data set is selected from [126] since it has less missing data than the data set of the arrest of a *cdc15* temperature-sensitive mutant, yet longer time-series than the data set of elutriation experiment. It consists of expression data of 6178 genes and 18 time points with 3.67% missing data. The protein-protein interaction data set is downloaded from DIP database [79] containing 18,272 interactions from 4,985 yeast proteins (as of Feb. 2008). Protein-protein interaction data stored in DIP database were obtained through manual curation of the scientific literature including both direct and complex interactions. Transcription factor binding data are from a data set consisting of the binding of almost all known yeast transcription factors monitored during cell growth in rich medium [60]. After excluding some probes for some computational reasons and problems with their microarray experiments, they provide binding data for 6229 genes across 203 transcription factors with 2.5% missing data.

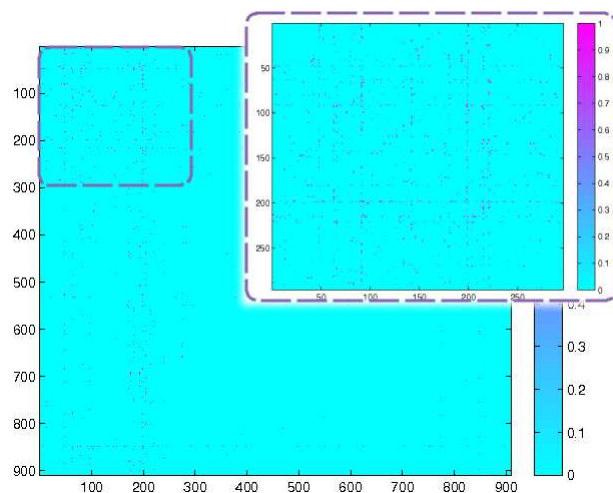
For the golden standard network/ground truth, we selected a yeast regulatory network from a comprehensive study [93]. The network was assembled from literature and a large amount of data, then divided into condition specific sub-networks including cell cycle, sporulation, diauxic shift, DNA damage and stress response. Altogether it contains 7,074 regulatory relationships between 142 transcription factors and 3,420 target genes. In this paper the cell cycle sub-network of 550 interactions among 296 genes is used as golden standard to compare with part of the resultant network.

To infer a cell-cycle specific network, we selected 909 genes by including the Spellman's 800 cell cycle's genes [126], Luscombe's 296 cell cycle's gene, and Price's 104 cell cycle genes [106]. Among these genes, there are 84 transcription factors. 2.7% of the gene expression data are missing for the 909 genes. There are 9 genes with 50% of their expression data missing and the corresponding data are discarded from use.

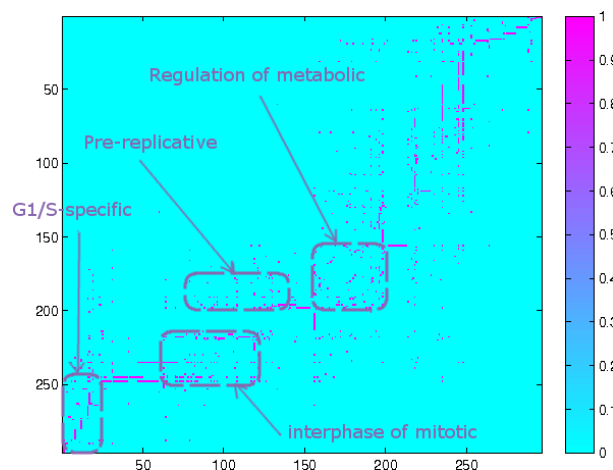
The rest of missing data are imputed by weighted K-nearest neighbours (KNNimpute) shown to be robust for microarray data [136]. For this gene set, the available transcription binding data are transcription factors binding 902 genes. 782 protein-protein interactions are found among all genes, which again composes of only a small fraction of all possible interactions (0.09%). All the data sets are available in the supplementary files.

BRFs inferred a network with 1,674 interactions between the 84 transcription factors and 669 genes, leaving 240 genes as irrelevant to the condition under study. The full adjacency matrix is shown in Figure 4.2(a). Since the network is too large to visualise, we select a sub-network of the 296 genes in the golden standard cell cycle network and plot it in Figure 4.3. The adjacency matrix of sub-network of 296 genes is illustrated in Figure 4.2(b) with four visible big clusters. This sub-network contains 608 interactions in total. Given the golden standard network, we can now investigate on the power of data integration. We address this issue by comparing the prediction power of individual data source and the integration result. By assuming there is a simple cut-off selection method for the coefficients, we plot the ROC curves for each data source in Figure 4.4. For example, since the binding data are the probabilities that a transcription factor binds to a gene, a cut-off threshold can be selected to include those interactions with higher probability than this threshold. Then the result of BRFs inference,  $\{p_l^1 | l = 1, 2, \dots, e\}$  instead of the binary matrix  $X$ , is plotted (red curve) in Figure 4.4.

Individual data source can only contribute to a weak predictor of the regulatory network, as can be seen from their ROC curves (black, violet and green) in Figure 4.4. This is consistent with previous findings [18, 91]. The predictive power with these data sources is often adversely affected by inherent factors of production techniques.



(a) The adjacency matrix of the inferred network, the part in dash frame corresponds to cell cycle specific sub-network.



(b) The modularized adjacency matrix of the cell cycle specific sub-network.

Figure 4.2: Experimental results for the 909 yeast genes.

For gene expression data, its ability to properly portray transcription is due to the experimental noise associated with the DNA microarray technique. As it is shown in Figure 4.4, the ROC curve (black) for gene expression partial correlation is indicative of its limited predictive power of the true network, although with a comprehensive coverage. Although binding is a necessary condition for regulatory activities, it may







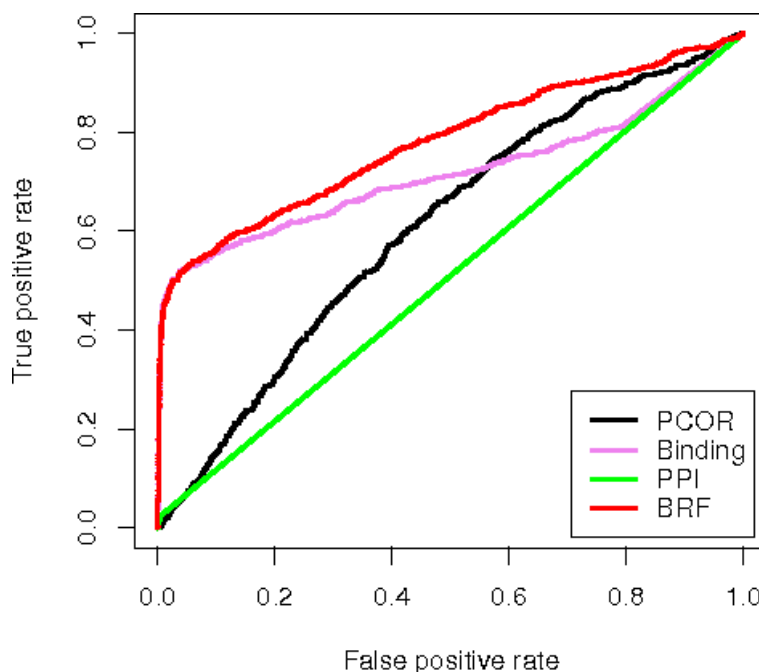


Figure 4.4: ROC curves for threshold selection methods for the three data types used in network reconstruction and the resultant network by BRFs on the 296-gene sub-network. PCOR stands for the partial correlation.

fraction of all possible interactions ( $296 \times 296 = 87,616$ ). In addition to the sparsity and poor quality, the main reason that protein-protein interaction alone achieves low predictive performance (green curve in Figure 4.4) lies in that less direct relationships exist between protein-protein interaction and the transcriptional network, since the protein-protein interaction data can only indicate potentials rather than presences of such interactions in the transcriptional process. This is also consistent with previous findings [11]. Nonetheless, the inclusion of a data set of low relevance and high noise into the integrative system reflects the robustness of the proposed algorithm, since the resultant network is neither biased to noise nor affected by the irrelevant information.

In Figure 4.5, the distribution of connectivity degree of nodes in the full network shows a power-law tail. To look for the functional modules in such a sparse network,

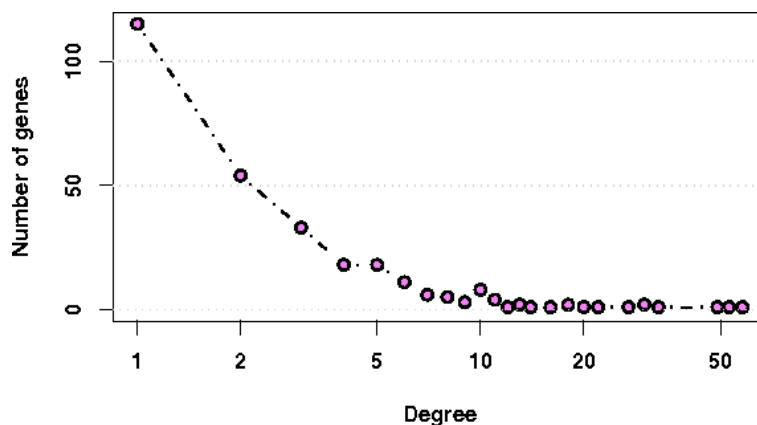


Figure 4.5: Connectivity degree distribution of the 296-gene sub-network. The  $x$  axis shows the degree of connectivity of 296 nodes on  $\log_2$  scale.

we study the transcriptional modules formed around hub-genes. The main hubs in this sub-network include transcriptional factors SWI4, SWI6, YOX1, MCM1, ACE2, etc. These 8 genes and their first neighbours cover 48% of 669 genes. We found that the clusters formed around these genes are significantly enriched with specific functions in Gene Ontology. The enrichment analysis result is provided in Table 4.2. Plots of time-series data of the eight transcription factors and the genes they are regulating are in Figure 4.6. Also we analysis the adjacency matrix of the 296-gene sub-network. Finding modules in gene networks is nontrivial, since the degree of overlapprotein-protein interactionng is high because of the existence of hubs. We focus on the four visible big clusters illustrated in Figure 4.2(b). Analysis on the function of genes within these clusters reveals 4 phase-specific modules as shown in Table 4.3.

Table 4.2: Over-represented GO terms in the transcriptional modules for eight transcription factors.

TF	GO term	<i>p</i> -values
SWI6	G1/S-specific transcription in mitotic cell cycle	6.68E-10
SWI6	regulation of cyclin-dependent protein kinase activity	2.14E-09
SWI6	regulation of cell cycle	4.02E-09
SWI6	mitotic cell cycle	3.53E-08
SWI6	regulation of kinase activity	6.71E-08
SWI4	biological regulation	2.41E-10
SWI4	G1/S-specific transcription in mitotic cell cycle	3.38E-10
SWI4	regulation of cellular process	4.79E-10
SWI4	interphase of mitotic cell cycle	5.50E-09
SWI4	regulation of cyclin-dependent protein kinase activity	8.87E-08
MBP1	regulation of cyclin-dependent protein kinase activity	7.47E-10
MBP1	regulation of kinase activity	2.37E-08
MBP1	mitotic cell cycle	3.24E-08
MBP1	mitotic sister chromatid cohesion	4.87E-07
MBP1	regulation of catalytic activity	8.09E-07
MCM1	mitotic cell cycle	3.36E-08
MCM1	biological regulation	7.40E-07
MCM1	interphase	1.22E-06
MCM1	regulation of cell cycle	1.82E-06
MCM1	pre-replicative complex formation	4.18E-05
FKH1	cell cycle	6.39E-07
FKH1	cell cycle phase	2.07E-06
FKH1	interphase of mitotic cell cycle	7.14E-06
FKH1	chromosome segregation	2.59E-05
FKH1	M phase of mitotic cell cycle	4.35E-05
SWI5	regulation of transcription from RNA polymerase II promoter by carbon catabolites	2.40E-05
SWI5	negative regulation of transcription from RNA polymerase II promoter by glucose	2.40E-05
SWI5	negative regulation of transcription	6.68E-05
SWI5	negative regulation of transcription from RNA polymerase II promoter	1.06E-04
SWI5	regulation of transcription, DNA-dependent	1.39E-04
YOX1	DNA replication	9.48E-05
YOX1	mitotic cell cycle	4.64E-04
YOX1	cell cycle process	1.07E-03
YOX1	regulation of cellular process	1.37E-03
YOX1	mitosis	1.81E-03
ACE2	regulation of transcription from RNA polymerase II promoter	9.58E-06
ACE2	regulation of biological process	2.74E-05
ACE2	regulation of transcription	3.52E-05
ACE2	negative regulation of transcription, DNA-dependent	1.85E-04
ACE2	regulation of cellular metabolic process	2.40E-04

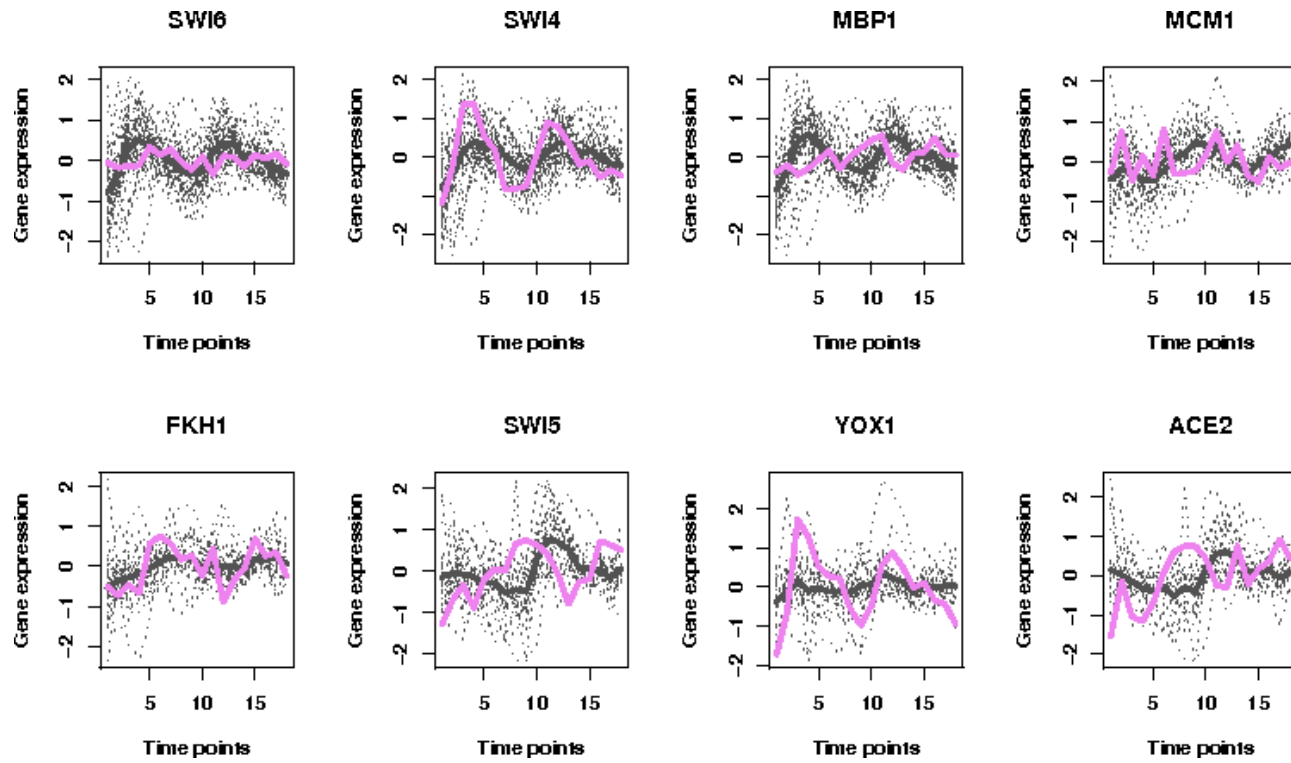


Figure 4.6: Plot of time series data of the eight transcription factors (pink) and the genes they are regulating (grey).

Table 4.3: Over-represented GO terms in four phases specific modules found in Figure 4.2(b).

Cluster	GO term	<i>p</i> -values	Gene counts
1	regulation of cellular metabolic process	6.14E-12	22
1	transcription, DNA-dependent	1.03E-11	22
1	regulation of transcription	2.40E-10	16
1	regulation of transcription, mating-type specific	8.82E-10	5
1	regulation of biological process	5.73E-09	21
2	regulation of transcription, mating-type specific	2.17E-05	3
2	transcription	5.80E-05	4
2	transcription, DNA-dependent	1.49E-04	13
2	regulation of biological process	1.73E-04	13
2	regulation of glycogen biosynthetic process	1.10E-03	2
3	biological regulation	9.67E-09	34
3	regulation of cyclin-dependent protein kinase activity	3.62E-08	6
3	regulation of catalytic activity	1.47E-07	8
3	regulation of kinase activity	1.08E-06	6
3	interphase of mitotic cell cycle	9.08E-05	8
4	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	5.31E-06	17
4	regulation of metabolic process	5.55E-06	19
4	DNA replication	9.18E-06	9
4	G1/S-specific transcription in mitotic cell cycle	1.05E-05	4
4	transcription	1.29E-05	19

## 4.5 Conclusions

Learning of large-scale regulatory networks is an important and challenging problem in bioinformatics. Although integrative analysis is promising in extracting deeper insights into the regulatory mechanism from diverse data sources, current methods are either bounded by the computational costs of microarray time-series analysis, or the difficulties of adapting to new data sources.

To address these issues, we proposed in this chapter a Bayes random fields method (BRFs) for integrative analysis of diverse genomic data. In the experiments on both synthetic and biological networks, BRFs shows superior performance. The success of BRFs is a direct result of the inherent elegant yet straightforward Bayesian integrative framework. Its flexibility enables unlimited heterogeneous data types to be integrated in a stochastic manner by a Gibbs sampler to facilitate robust estimation. As previously addressed, different data are of various formats and sparsity. BRFs propagates through modelling the two distributions in the available data sets without resorting to accounting for missing data, thus is more effective. In particular, the random fields component introduces a known feature of gene network for accurate modelling.

We are aware of the limitation of graphical Gaussian models/partial correlation that in theory it is not as powerful as dynamic Bayesian networks approaches when there are non-linear effects present in the data. However, given the paucity of samples available and the large scale of network, it is impossible for full order Bayesian inference with time-series data.

# Chapter 5

## Conclusions and Future Research

### 5.1 Conclusions

This thesis is focused on applying innovative statistical inference methodologies to genomics research supported by high-throughput biotechnologies. This multidisciplinary research area facilitates progress in not just computer science, but also statistics, and molecular biology. Therefore, the main thread of this thesis is that applications developed for genomics research and statistical inference techniques can be synergistic and pursue advancements together.

The main contributions of this thesis are summarised as follows.

- Proposing a tight clustering method based on partial mixture model to address the need of obtaining tighter and therefore more biologically meaningful gene clusters.
- Proposing a GO-driven clustering validation index that not only makes full use of GO annotation information but also systematically takes GO's structure information into account.
- Proposing a gene regulatory network inference method to make integrative in-

ference from multiple data sources with increased predictive accuracy.

In their individual research fields, these proposed methods bridge research gaps in the literature, as elaborated in the following sections, Section 5.1.1 to 5.1.3.

### 5.1.1 Partial Mixture Model Tight Clustering

In Chapter 2, tight clustering was achieved based on the flexibility of partial mixture model and the robustness of the minimum distance estimator. Previously, partial mixture model was known to be capable of solving problems for low dimensional data. In fact, one problem with the classical partial mixture model is that it cannot fit data of more than 7 data points [117]. In Chapter 2, a partial mixture model was extended to be used on high dimensional data by integrating it with a spline regression model. By partial modelling, the mixture model is allowed to find the core component in the data. On the other hand, the unsupervised manner of the proposed method in its selection of cluster number makes it a powerful tool for clustering gene expression data.

In the experiments reported in Chapter 2, the proposed algorithm was validated and its clustering outcomes including both gene clusters and scattered genes were explained with the help of various biological resources. Because of the incomplete biological knowledge, no conclusion can be drawn by merely comparing clustering results with known measures from the biological literature. Therefore, besides using a data-driven index, GO enrichment analysis was applied to the clustering outcomes. From current knowledge, it is proved that the proposed clustering method can help separate groups of genes with similar functions, while new hypothesis can be obtained by exploring the scattered genes. Findings from this study have been published in two papers [155, 157].



As a result, this chapter provides an excellent example of gene expression data mining by combining machine learning techniques and biological knowledge. Further, the tight clustering method can be applied to other areas that require outlier detection and tightly correlated data points, such as neural signal processing and image segmentation. Gene annotations help reveal new hypothesis derived from the inference of the scattered genes. One concern about the Gene Ontology analysis and gene annotations is that many genes and their functions are still unknown or poorly understood. It is our hope that through clustering, new understanding about these genes can be introduced to genomics research. It is also this ambition that inspires a GO-driven validation method proposed in Chapter 3.

### 5.1.2 Clustering Validation Using Gene Ontology

In Chapter 3, a literature review revealed that existing GO-driven validation methods either fail to achieve power when facing the redundant and complex structure of GO, or tend to ignore the intrinsic properties of GO categories, as the experimental and analytical results reported respectively in Section 3.2 and 3.3. Assessing clustering quality with existing biological knowledge that is manually curated into biological databases is a promising direction, however, special attention has to be paid not only because of the complex structure and unique features of GO, but also because the biological information in GO is still noisy, incomplete and sometimes even erroneous.

To take the structure information in GO into account and make full use of GO annotations, two clustering validation indices and a combined index which apply graph theory and theory of hypothesis testing were proposed in Chapter 3. In particular, the Functional Compactness measure and Functional Similarity measure can be used for

evaluating how closely genes within a cluster are related to each other and finding commonality between two clusters. These indices not only take into account the intrinsic properties of GO categories, but also integrate information from GO's graphical structure.

The proposed GO-driven indices achieve robustness to noise in GO using a pooling technique, as experimentally proven in the comparative experiments. In these experiments that are designed to test various validity indices' consistency, accuracy and discriminability, the proposed WB index demonstrated superior performance throughout. In summary, this chapter provides excellent examples of exploiting GO information to facilitate integrative analysis of experimental results and existing knowledge, and further providing quantitative supports for validation studies. As a result, the method has been published as a conference paper [158], and a journal paper is currently in peer review.

Statistical clustering is an active research area, which has seen many advances in recent years. In contrast, biological evidences on even the most well-studied organism are accumulated and organised into databases only since recently. Therefore, it is preferable to process experimental data sets with pure statistic methods and systematically compare results with known biological facts for new knowledge discovery. Indeed, it is the contradiction between statistical findings and current biological knowledge that stimulates interests and propels developments in the post-genome era. Results from biological validation, especially for less annotated and higher eukaryotic organism, need to be carefully analysed and interpreted, such as the case of *Arabidopsis Thaliana* in Section 3.5.1.2.

### 5.1.3 Transcriptional Regulatory Network Reconstruction

To address the many issues in reconstructing transcriptional regulatory networks, as raised in Section 1.2.3, a Bayes Random Fields approach for transcriptional regulatory network reconstruction was presented in Chapter 4. With a Gibbs sampler, the approach enjoys rigorous inference and robust analysis from large-scale genomic data whilst minimising the influence of inherent noise.

The proposed method's flexibility benefits from a full Bayesian routine which enables integrative analysis of a wide range of data sources. It can be easily adapted to new data sources, yet remains efficient enough to facilitate large-scale analysis for discovering relevant network architecture. This is achieved by first providing inference results for the high-dimensional gene expression data to be included into the integrative system. The time series inference method is selected based on its experimentally proven suitability for microarray data. Moreover, the tight clustering method proposed in Chapter 2 is integrated in the random fields component to impose a modular constraint on the resultant network, so as to introduce a modularity feature of biological networks.

In comparison to the many works on extensive network reconstruction for the whole organism, the proposed probabilistic network inference approach integrates evidence from a diversity of resources in a seamless and coherent manner. It identifies network scaffold on a context specific level, as it is shown in the experiment to reveal cell-cycle relevant subnetwork. As a result, this work has been published as a journal paper [156].

## **5.2 Future Research**

Significant breakthroughs in biotechnology in recent years have the potential of bringing the genomics research to public health care. For example, low cost sequencing tools may enable genetic tests on a clinical level, which are revolutionary not only in discovery of genetic disposition to a certain disease, but also in making personalised medicine possible [42]. Drugs and drug combinations designed with respect to the patients' genotype can be optimised to ensure maximum efficacy with minimal adverse effects.

The major obstacle in this field is still the incomplete understanding in functional genomics. While the rapid development of high-throughput biotechnology is making large amount of genomic data available, there is a lack of software and genomic knowledge for effective data analysis. Enormous amounts of the resulting data from high-throughput biotechnology have drawn attentions in the bioinformatics community. Innovative and objective inference methods are urgently required in genomics research to reveal the mechanism underneath complex biological systems.

Looking forward, on the basis of current research progress in bioinformatics, continuous efforts are needed in designing computational techniques and developing analytical software to help consolidate the foundation of functional genomics. Among the many features these techniques should possess, robustness to noise in genomic data and flexibility in the inference procedure are the key. In the near future, the following research proposals can be considered.

### 5.2.1 Inferring Causal Relations from Large-scale Gene Expression Data

During the reconstruction of large-scale gene regulatory networks, as described in Chapter 4, inefficiency of even some of the most popular methods in time series inference is found. Identifying regulatory relationships between genes based on their expression time series is essentially equivalent to quantifying causal relations between time series. Therefore, many existing methods for learning causal relations have been adapted to this field. Some methods use Granger causality [55], a statistical technique for causal inference well known in economics. [3] provides a good review on current network reconstruction methods based on gene expression data. However, simple adaptations of existing time series inference methods rarely succeed, both because of the small sample size and the large number of variables in microarray gene expression time series.

Moreover, resulting large-scale network from the BRFs method proposed in Chapter 4 needs to be further refined into a directed form, so that its regulatory mechanism can be revealed. Combining these two issues, current directed time series inference methods should be evaluated to observe their performance on microarray data or to discover the intrinsic problems that causes their inefficiency. In this way, new method targeted on these problems may be found and tested. As the first step, inference techniques such as graphical Gaussian models and dynamic Bayesian networks need to be implemented and tested. When facing the scale of microarray gene expression data, efficiency is the key to a successful analysis tool.

### **5.2.2 GO-driven Validity Index for Regulatory Network Inference Methods**

Biological annotation data can be of great help not only in gene clustering validations, but also in validating gene network inference results. From February 2009, regulatory relationships between GO terms will be implemented in GO, which make it potentially promising for new, quantitative validity index for regulatory network inference methods. Although, in some cases, the accuracy and completeness of gene annotation is by no means sufficient, these annotation can serve as useful prior knowledge for validation. This can be a new line of investigations in the near future. However, there are two issues that researchers should take into account.

First, the regulatory relationships will be implemented in the BP ontology, the MF ontology, and between the BP and MF ontologies. This means these two ontologies are no longer strictly independent. Second, regulatory relationships will be presented as three types of relationships: ‘regulates’, ‘positively\_regulates’ and ‘negatively\_regulates’ relationships. They provide descriptions for interactions between biological processes, molecular functions or biological qualities. While the addition of these relationships improves the ability of the ontology to represent biology completely and accurately, an implication of the change is that future tools can no longer ignore GO relationship type. The tools also must be compatible with inter-ontology links between GO categories.

### **5.2.3 Combined Analysis of DNA Sequence and Microarray Data**

By jointly modelling diverse types of data, additional insight into complex biology systems may be gained. On the basis of the BRFs integrative framework and the partial

mixture model clustering method, sequence and gene expression data can be jointly modelled to form gene clusters that simultaneously maximise feature cardinality from both data sources.

In this way, behaviour changes in gene expression can be explained by common regulatory mechanisms at the transcriptional level. Also, the results from combinatory analysis will not only help identify combinatorial regulation relationships among genes, but also provide insights into the regulatory mechanism for individual genes. A promising direction is to use reliable multi-objective optimisation techniques in machine learning. From the machine learning point of view, this problem is essentially a multi-objective optimisation problem. A solution should allow one to counterbalance the bias from different objectives through the simultaneous maximisation of feature cardinality.

To conclude, a fundamental problem for applying computational methodologies to functional genomics is, that the attributes of genomic data do not fit the assumptions classical inference techniques often make. Objective inference methods are then needed, both to design suitable models for genomic data and to provide robust inference procedure against biological noise. Meanwhile, high-throughput technologies supporting functional genomics is rapidly evolving. Advancements in biotechnologies require innovative and powerful analytical software to meet new demands on a regular basis.

Therefore, it is essential for researchers to keep up with the fast pace and design new inference methods to fulfill the requirements from genomic data. Methods should not only provide robustness to noise, flexibility in capturing arbitrary, overlapping and agglomerative attributes of the genomic data, but also remain computationally efficient enough for the large-scale nature of data. The lack of biological ground truth even

in the most well-studied organism means that results gained from computational analysis tools need to be interpreted carefully in order to draw reliable conclusions and validations. Meanwhile, just as inspirations for this field can be gained from other disciplines, insights gained from research on complex biological systems may in turn contribute to applications in other scientific fields.



# Appendix A

## A.1 Theoretical Comparison between MDE and MLE

Both minimum distance estimator and maximum likelihood estimator belong to the Minimum distance (MD) family. Given the parameters vector of interest  $\theta_0 \in \Theta$  where  $\Theta$  is the set of possible parameter values, the aim of MD estimators can be generalised as the minimisation of a criterion function

$$F(\theta) = \hat{g}(\theta)D_w(\theta), \tag{A.1}$$

where  $\hat{g}(\theta)$  is a function of the data  $y_t$  that will verify  $\hat{g}(\theta_0) \rightarrow 0$ , and  $D_w(\theta)$  being a weighted distance matrix. Depending on the choice of  $\hat{g}(\theta)$ , different estimators can be generated.

In particular, a minimum divergence estimator, which incorporates minimum distance and maximum likelihood, is proposed [7] as an alternative to non-parametric density estimation. Density-based minimum divergence methods include those estimate parameters through minimising some pre-defined divergence between the assumed model density and the true model density underlying the data, e.g. maximum

likelihood method and minimum chi-squared method. The criterion is given by

$$\theta = \arg \min_{\theta} \left[ \int f(x|\theta)^{1+\alpha} dx - \frac{1+\alpha}{n\alpha} \sum_{i=1}^n f(x_i|\theta)^{\alpha} \right], \quad (\text{A.2})$$

with a metaparameter  $\alpha > 0$ . MDE corresponds to  $\alpha = 1$  while MLE corresponds to  $\alpha \rightarrow 0$ .

Examples of the two estimation criteria for normal density  $X \sim \mathcal{N}(\mu, \sigma^2)$  are

$$\hat{\mu}_{MLE} = \arg \max_{\mu} \sum_{i=1}^n \log \phi(x_i|\mu, \sigma^2), \quad (\text{A.3})$$

$$\hat{\mu}_{MDE} = \arg \min_{\mu} \left( \frac{1}{2\sigma\sqrt{\pi}} - \frac{2}{n} \sum_{i=1}^n \phi(x_i|\mu, \sigma^2) \right). \quad (\text{A.4})$$

While the aim of MDE is to maximise the sum of the densities, MLE tries to maximise the product of the densities.

## A.2 Mean Integrated Squared Error

For analysing more than one dataset, the Mean Integrated Squared Error (MISE) is a more appropriate error criteria for the kernel density estimator [112]. Let  $\hat{f}(x)$  be an estimator of the density function  $f(x)$  given  $n$  samples  $x_i, i = 1, 2, \dots, n$ ,

$$MISE(\hat{f}(x)) = E \int [\hat{f}(x) - f(x)]^2 dx, \quad (\text{A.5})$$

which, by changing the order of integration, is the integral of the mean squared error (MSE):

$$MSE(\hat{f}) = E(\hat{f} - f)^2 = Var(\hat{f}) + (E\hat{f} - f)^2. \quad (A.6)$$

Suppose  $\kappa$  satisfying  $\int \kappa(x)dx = 1$  is the kernel for the kernel density estimator,

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n \kappa_h(x - x_i), \quad (A.7)$$

where  $\kappa_h = 1/h\kappa(u/h)$ ,  $h$  being the bandwidth. From [143, Section 2.3],

$$MSE(\hat{f}) = \frac{1}{n} [(\kappa_h^2 * f)(x) - (\kappa_h * f)^2(x)] + [(\kappa_h * f)(x) - f(x)]^2, \quad (A.8)$$

with the convolution notation

$$(f * g)(x) = \int f(x - y)g(y)dy. \quad (A.9)$$

These may be combined to give

$$MISE(\hat{f}) = \frac{1}{nh} \int \kappa(x)^2 dx + (1 - \frac{1}{n}) \int (\kappa_h * f)^2(x) dx - 2 \int (\kappa_h * f)(x) f(x) dx + \int f(x)^2 dx. \quad (A.10)$$

Thus by using Eq.(A.8), exact MISE expressions can be derived. For the Gaussian mixture densities in Eq.(2.13), MISE has the form:

$$\begin{aligned} MISE(\hat{f}) \\ = \frac{1}{2nh\sqrt{\pi}} + W^T [(1 - \frac{1}{n})\Omega_2 - 2\Omega_1 + \Omega_0] W, \end{aligned} \quad (A.11)$$

where  $W$  is the vector for weight parameters,  $\Omega_a$  is a  $K \times K$  matrix with the element  $(i, j)$  corresponding to  $\phi^{(ah^2 + \sigma_i^2 + \sigma_j^2)^{1/2}}(\mu_i - \mu_j)$ . Eq.(A.11) entitles a rich family of Gaussian mixture models to be estimated. The first item in Eq.(A.11) does not change when minimising  $MISE$ . Therefore we obtain a new criterion for model fitting with respect to mean integrated squared error

$$\begin{aligned}\theta &= \arg \min_{\theta} MISE(\hat{f}) \\ &= \arg \min_{\theta} \{W^T [(1 - \frac{1}{n})\Omega_2 - 2\Omega_1 + \Omega_0]W\}.\end{aligned}\tag{A.12}$$

# Appendix B

By the standard graphical theory, an efficient way of obtaining partial correlation matrix is through the inverse of covariance or correlation matrix [12]. However, classical time series analysis techniques are not readily applicable to transcriptomic data, in which the number of data points  $n$  far exceeds the sample size  $t$ , since in this case the sample covariance and correlation matrices are not positively definite. Recently, an efficient way for computing partial correlation was proposed by using only the  $t - 1$  eigenvectors corresponding to the  $t - 1$  non-zero eigenvalues of the covariance matrix [83]. Such a dimension reconstruction method is popular in signal processing community and known to be robust against noise.

## B.1 Efficient Computation of Partial Correlation

This section describes an efficient computation method of partial correlation when the number of data points  $n$  far exceed the sample size  $t$ , i.e.  $n \gg t$ . In the BRFs framework, it is proposed to use partial correlation metric as the inference result from time series data.

By removing the linear effects from the rest of population, partial correlation can

indicate whether a pair of variables directly interact with each other. Because of its efficiency, partial correlation has been the foundation for graphical Gaussian models. The aim is to set up a graphical interaction model  $G = (V, E)$  with the vertices  $\{V\}$  as the components of the series and edges  $\{E\}$  denoting pair-wise interactions. graph  $G$  have such property

$$E(i, j) \subseteq G \Leftrightarrow y_i \perp\!\!\!\perp y_j | Y_{-ij}. \quad (\text{B.1})$$

Let  $Y_{-ij} = \{y_k | k \neq i, j\}$ , the linear effects of  $Y_{-ij}$  is removed from  $y_i$  by finding the parameter set  $\theta_i = (\mu_i, \kappa_i)$  such that

$$\hat{\theta} = \arg \min_{\theta} E \left\{ y_i(t) - \mu_i - \sum_u \kappa_i(t-u) Y_{-ij}(u) \right\}^2. \quad (\text{B.2})$$

The residuals of such regression is denoted as  $\epsilon_i$ . In the same way we define  $\epsilon_j$ . Thus the correlation between residuals  $\epsilon_i$  and  $\epsilon_j$  is the correlation between variables  $y_i$  and  $y_j$  conditioned on the others, i.e., partial correlation between  $y_i$  and  $y_j$ . A direct interaction between  $y_i$  and  $y_j$  exists if and only if their partial correlation is significantly different from zero. When partial correlation is applied for network reconstruction, it provides solid mathematical foundation for finding meaningful interactions. It leads to the definition of the graph

$$E(i, j) \subseteq G \Leftrightarrow \text{cor}(\epsilon_i, \epsilon_j) = 0. \quad (\text{B.3})$$

An efficient way of obtaining partial correlation matrix, by the standard graphical theory, is through the inverse of covariance or correlation matrix [12]. Based on this theory, partial spectral coherence was proposed for frequency domain analysis of time series [12] and it can be obtained by the inverse of the spectral matrix [28]. However,

these classical time series analysis techniques are not readily applicable to transcriptome data, where the number of data points  $n$  far exceed the sample size  $t$ , i.e.  $n \gg t$ . Since in this case the sample covariance and correlation matrices are not positively definite. Many efforts were spent on exploiting this field, either by restricting inference to a small number of genes [145], or limiting partial coefficient to limited order [34, 149], i.e., computation is conditioned on only limited number of genes each time. Sampling technique such as bootstrapping is also proposed [113] in order to obtain point estimates of partial correlation coefficient. Recently, it was proved that the partial correlation matrix maximises the entropy of interaction system [83], and an efficient computation of partial correlation was proposed by using only the  $t - 1$  eigenvectors corresponding to the  $t - 1$  non-zero eigenvalues [83]. Such reconstruction method is popular in signal processing community and known to be robust against small noise. Let

$$V = \{v \in V, Cv = \lambda v\} \quad (\text{B.4})$$

be the eigenvector of  $C$ , and  $\lambda_i, i = 1, \dots, N$  be the eigenvalues. Since the spectral decomposition of covariance matrix  $C$  is

$$C = M\Lambda M^{-1}, \quad \{\Lambda_{ii}\} = \lambda_i. \quad (\text{B.5})$$

There are exactly  $t - 1$  non-zero eigenvalues, partial correlation matrix can be constructed in the non-zero eigenspace

$$P = C^{-1} = (M\Lambda M^{-1})^{-1} = M\Lambda^{-1}M^{-1}. \quad (\text{B.6})$$

$M$  is a matrix whose columns are made up of eigenvectors  $v$ , and  $\Lambda$  is a diagonal

## B.1 Efficient Computation of Partial Correlation

---

matrix whose diagonal elements are the corresponding eigenvalues  $\lambda$ , therefore  $\hat{P}$  can be reconstructed using the  $t - 1$  eigenvectors corresponding to  $\{\lambda_1, \lambda_2, \dots, \lambda_{t-1}\}$ .



# References

- [1] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.
- [2] Azuaje, F. and Bodenreider, O. (2004). Incorporating ontology-driven similarity knowledge into functional genomics: An exploratory study. In *Proceedings of the 4th IEEE Symposium on Bioinformatics and Bioengineering*, pages 317– 324.
- [3] Bansal, M., Belcastro, V., Ambesi-Impiombato, A., and di Bernardo, D. (2007). How to infer gene networks from expression profiles. *Molecular Systems Biology*, **3**, 78.
- [4] Bar-Joseph, Z., Gerber, G., Gifford, D. K., Jaakkola, T. S., and Simon, I. (2002). A new approach to analyzing gene expression time series data. *Proceedings of the Annual International Conference on Computational Molecular Biology, RECOMB*, pages 39–48.
- [5] Bar-Joseph, Z., Gerber, G. K., Lee, T. I., Rinaldi, N. J., Yoo, J. Y., Robert, F., Gordon, D. B., Fraenkel, E., Jaakkola, T. S., Young, R. A., and Gifford, D. K. (2003). Computational discovery of gene modules and regulatory networks. *Nature Biotechnology*, **21**(11), 1337–1342.

- [6] Barriot, R., Sherman, D., and Dutour, I. (2007). How to decide which are the most pertinent overly-represented features during gene set enrichment analysis. *BMC Bioinformatics*, **8**(1), 332.
- [7] Basu, A., Harris, I., Hjort, N., and Jones, M. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, **85**, 549–559.
- [8] Ben-Hur, A. and Noble, W. (2004). Pre-mRNA splicing: life at the centre of the central dogma. *Journal of Cell Science*, (117), 6261–3.
- [9] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, **B**(57), 289–300.
- [10] Beran, R. (1984). Minimum distance procedures. *Handbook of Statistics*, **4**, 741–754.
- [11] Bernard, A. and Hartemink, A. J. (2005). Informative structure priors: joint learning of dynamic regulatory networks from multiple types of data. *Proceedings of the Pacific Symposium on Biocomputing*, pages 459–70.
- [12] Billinger, D. (1981). *Time series: Data Analysis and Theory*. Holt, Rinehart and Winston, New York.
- [13] Bolshakova, N. and Azuaje, F. (2003). Cluster validation techniques for genome expression data. *Signal Processing*, pages 825–833.
- [14] Bolshakova, N., Zamolotskikh, A., and Cunningham, P. (2006a). Comparison of the data-based and gene ontology-based approaches to cluster validation methods

- for gene microarrays. In *Proceedings of International Symposium on Computer-Based Medical Systems*, pages 539–543.
- [15] Bolshakova, N., Azuaje, F., and Cunningham, P. (2006b). Incorporating biological domain knowledge into cluster validity assessment. In *Applications of Evolutionary Computing*, pages 13–22.
- [16] Boulesteix, A. L. and Strimmer, K. (2007). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Brief Bioinformatics*, **8**(1), 32–44.
- [17] Boutros, P. C. and Okey, A. B. (2005). Unsupervised pattern recognition: An introduction to the whys and wherefores of clustering microarray data. *Brief Bioinformatics*, **6**(4), 331–343.
- [18] Brynildsen, M. P., Tran, L. M., and Liao, J. C. (2006). A Gibbs sampler for the identification of gene expression and network connectivity consistency. *Bioinformatics*, **22**(24), 3040–3046.
- [19] Bussemaker, H., Li, H., and Siggia, E. (2001). Regulatory element detection using correlation with expression. *Nature Genetics*, **27**(2), 167–71.
- [20] Butte, A. J. and Kohane, I. S. (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Proceedings of the Pacific Symposium on Biocomputing*, **5**, 415–426.
- [21] Calinski, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in statistics*, **3**, 1–27.

- [22] Cheng, J., Martin, J., Cline, M., Awad, T., and Siani-Rose, M. (2002). Gene expression profiling analysis augmented by mathematically transformed gene ontology. *Proceedings of International Conference on Intelligent Systems in Molecular Biology*.
- [23] Cherry, J. M., Adler, C., Ball, C., Chervitz, S. A., Dwight, S. S., Hester, E. T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., Weng, S., and Botstein, D. (1998). Sgd: Saccharomyces genome database. *Nucleic Acids Research*, **26**(1), 73–79.
- [24] Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J., and Davis, R. W. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell*, **2**(1), 65–73.
- [25] Clarke, R., Ransom, H. W., Wang, A., Xuan, J., Liu, M. C., Gehan, E. A., and Wang, Y. (2008). The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nature Revolution Cancer*, **8**(1), 37–49.
- [26] Couto, F., Silva, M., and Coutinho, P. (2003). Implementation of a functional semantic similarity measure between gene-products. DI/FCUL TR 03–29, Department of Informatics, University of Lisbon. November.
- [27] Crick, F. (1970). Central dogma of molecular biology. *Nature*, **227**(258), 561C3.
- [28] Dahlhaus, R. (1999). Graphical interaction models for multivariate time series. *Metrika*, **51**, 157–172.
- [29] Datta, S. and Datta, S. (2006a). Evaluation of clustering algorithms for gene expression data. *BMC Bioinformatics*, **7**(1), S17. Dec.

- [30] Datta, S. and Datta, S. (2006b). Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. *BMC Bioinformatics*, **7**(1), 397.
- [31] Davidon, W. C. (1991). Variable metric method for minimization. *SIOPT*, **1**, 1–17.
- [32] Davidson, E. H., Rast, J. P., Oliveri, P., Ransick, A., Calestani, C., Yuh, C.-H., Minokawa, T., Amore, G., Hinman, V., Arenas-Mena, C., Otim, O., Brown, C. T., Livi, C. B., Lee, P. Y., Revilla, R., Rust, A. G., jun Pan, Z., Schilstra, M. J., Clarke, P. J. C., Arnone, M. I., Rowen, L., Cameron, R. A., McClay, D. R., Hood, L., and Bolouri, H. (2002). A Genomic Regulatory Network for Development. *Science*, **295**(5560), 1669–1678.
- [33] Davies, J. and Bouldin, D. (1979). A cluster separation measure. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 1, pages 224–227.
- [34] de la Fuente, A., Bing, N., Hoeschele, I., and Mendes, P. (2004). Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*, **20**(18), 3565–3574.
- [35] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, **39**(Series B), 1–38.
- [36] Dennis, J. E. and Schnabel, R. B. (1983). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall.

- [37] Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, **1**, 269–271.
- [38] Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G., and West, M. (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, **90**(1), 196–212.
- [39] Dojer, N., Gambin, A., Mizera, A., Wilczynski, B., and Tiuryn, J. (2006). Applying dynamic Bayesian networks to perturbed gene expression data. *BMC Bioinformatics*, **7**, 249.
- [40] Dougherty, E. R. and Datta, A. (2005). Genomic signal processing: Diagnosis and therapy. *IEEE Signal Processing Magazine*, **22**(1), 107–112.
- [41] Dunn, J. (1974). Well separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, **4**, 95–104.
- [42] Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., deWinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., and Turner, S. (2008). Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*, pages 133–138.
- [43] Ernst, J., Nau, G. J., and Bar-Joseph, Z. (2005). Clustering short time series gene expression data. *Bioinformatics*, **21**(SUPPL. 1).

- [44] Fraley, C. and Raftery, A. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, **97**(458), 611–631.
- [45] Fraley, C. and Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal*, **41**(8), 578–588.
- [46] Fraley, C. and Raftery, A. E. (2003). Enhanced model-based clustering, density estimation, and discriminant analysis software: Mclust. *Journal of Classification*, **20**(2), 263–286.
- [47] Fraley, C. and Raftery, A. E. (2006). Mclust version 3: an R package for normal mixture modeling and model-based clustering. *Technical Report 504, Department of Statistics, University of Washington, Seattle*.
- [48] Friedman, J. H. and Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modeling. *Technometrics*, **31**, 35.
- [49] Friedman, N. (2004). Inferring Cellular Networks Using Probabilistic Graphical Models. *Science*, **303**(5659), 799–805.
- [50] Friedman, N., Linial, M., Nachman, I., and Pe’er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, **7**(3), 601–620.
- [51] Gao, F., Foat, B., and Bussemaker, H. (2004). Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics*, **5**(1), 31.

- [52] Garcia-Dorado, A. and Gallego, A. (2003). Comparing Analysis Methods for Mutation-Accumulation Data: A Simulation Study. *Genetics*, **164**(2), 807–819.
- [53] Gill, J. (2002). *Bayesian methods : a social and behavioral sciences approach*. Chapman & Hall/CRC, Boca Raton, Fla.
- [54] Giudici, P. and Green, P. (1999). Decomposable graphical Gaussian model determination. *Biometrika*, **86**, 785–801.
- [55] Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, **37**, 424–438.
- [56] Guelzim, N., Bottani, S., Bourguin, P., and Képès, F. (2002). Topological and causal structure of the yeast transcriptional regulatory network. *Nature Genetics*, **31**(1), 60 – 63.
- [57] Guo, X., Liu, R., Shriver, C. D., Hu, H., and Liebman, M. N. (2006). Assessing semantic similarity measures for the characterization of human regulatory pathways. *Bioinformatics*, **22**(8), 967–973.
- [58] Halliday, D. (2000). Temporal correlation of large scale synaptic input is a major determinant of neuronal bandwidth. *Neural Computation*, **12**, 693–707.
- [59] Hanisch, D., Zien, A., Zimmer, R., and Lengauer, T. (2002). Co-clustering of biological networks and gene expression data. In *Proceedings of International Conference on Intelligent Systems for Molecular Biology*, pages 145–154.
- [60] Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., Hannett, N. M., Tagne, J.-B., Reynolds, D. B., Yoo, J., Jennings, E. G.,



- Zeitlinger, J., Pokholok, D. K., Kellis, M., Rolfe, P. A., Takusagawa, K. T., Lander, E. S., Gifford, D. K., Fraenkel, E., and Young, R. A. (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99 – 104.
- [61] Hartemink, A. J., Gifford, D. K., Jaakkola, T. S., and Young, R. A. (2002). Combining location and expression data for principled discovery of genetic regulatory network models. In *Proceedings of Pacific Symposium on Biocomputing* 7:437-449.
- [62] Heard, N. A., Holmes, C. C., Stephens, D. A., Hand, D. J., and Dimopoulos, G. (2005). Bayesian coclustering of anopheles gene expression time series: Study of immune defense response to multiple experimental challenges. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(47), 16939–16944.
- [63] Heard, N. A., Holmes, C. C., and Stephens, D. A. (2006). A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes: An application of Bayesian hierarchical clustering of curves. *Journal of the American Statistical Association*, **101**(473), 18–29.
- [64] Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, **2**, 193–218.
- [65] Husmeier, D. (2003). Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics*, **19**(17), 2271–2282.
- [66] Huttenlocher, D. P., Klanderman, G. A., and Rucklidge, W. A. (1993). Comparing images using the Hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **15**(9), 850–863.

- [67] Ideker, T., Thorsson, V., Ranish, J., Christmas, R., Buhler, J., Eng, J., Bumgarner, R., Goodlett, D., Aebersold, R., and Hood, L. (2001). Integrated genomic and proteomic analyses of a systemically perturbed metabolic network. *Science*, **292**, 929–934.
- [68] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America*, **98**(8), 4569–4574.
- [69] J. Jiang, D. C. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *Proceedings on International Conference on Research in Computational Linguistics*, pages 19–33.
- [70] Jeong, H., Mason, S. P., Barabasi, A.-L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, **411**, 41.
- [71] Ji, H. and Wong, W. H. (2006). Computational biology: Toward deciphering gene regulatory information in mammalian genomes. *Biometrics*, **62**, 645–663(19).
- [72] Jiang, D., Tang, C., and Zhang, A. (2004a). Cluster analysis for gene expression data: a survey. *IEEE Transactions on Knowledge and Data Engineering*, **16**(11), 1370–1386.
- [73] Jiang, D., Pei, J., Ramanathan, M., Tang, C., and Zhang, A. (2004b). Mining coherent gene clusters from gene-sample-time microarray data. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 430–439, New York, NY, USA. ACM Press.

- [74] Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., and West, M. (2005). Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science*, **20**, 388–400.
- [75] Joyce, A. R. and Bernhard (2006). The model organism as a system: integrating ‘omics’ data sets. *Nature Reviews Molecular Cell Biology*, **7**(3), 198–210.
- [76] Kato, M., Hata, N., Banerjee, N., Futcher, B., and Zhang, M. Q. (2004). Identifying combinatorial regulation of transcription factors and binding motifs. *Genome Biology*, **5**(8).
- [77] Kaufman, L. and Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley and Sons, New York.
- [78] Kay, S. M. (1993). *Fundamentals of statistical signal processing: estimation theory*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.
- [79] L, S., CS, M., AJ, S., FK, P., JU, B., and D, E. (2004). The database of interacting proteins: 2004 update. *Nucleic Acids Research*, **32**, 449–451.
- [80] Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- [81] Lebre, S. (2007). Inferring dynamic genetic networks with low order independencies.
- [82] Lee, I., Date, S. V., Adai, A. T., and Marcotte, E. M. (2004). A Probabilistic Functional Network of Yeast Genes. *Science*, **306**(5701), 1555–1558.

- [83] Lezon, T. R., Banavar, J. R., Cieplak, M., Maritan, A., and Fedoroff, N. V. (2006). Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*.
- [84] Li, C.-T., Yuan, Y., and Wilson, R. (2008). An unsupervised conditional random fields approach for clustering gene expression time series. *Bioinformatics*, **24**(21), 2467–2473. November.
- [85] Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the International Conference on Machine Learning*, pages 296–304.
- [86] Liu, X., Jessen, W. J., Sivaganesan, S., Aronow, B. J., and Medvedovic, M. (2007). Bayesian hierarchical model for transcriptional module discovery by jointly modeling gene expression and chip-chip data. *BMC Bioinformatics*, **8**, 283+. August.
- [87] Locke, J. C. W., Millar, A. J., and Turner, M. S. (2005). Modelling genetic networks with noisy and varied experimental data: the circadian clock in *arabidopsis thaliana*. *Journal of Theoretical Biology*, **234**, 383–393.
- [88] Loganantharaj, R. and Atwi, M. (2007). Towards validating the hypothesis of phylogenetic profiling. *BMC Bioinformatics*, **8**, 25.
- [89] Loganantharaj, R., Cheepala, S., and Clifford, J. (2006). Metric for measuring the effectiveness of clustering of DNA microarray expression. *BMC Bioinformatics*, **7**, S5.

- [90] Lord, P. W., Stevens, R. D., Brass, A., and Goble, C. A. (2003). Semantic similarity measures as tools for exploring the gene ontology. *Proceedings of the Pacific Symposium on Biocomputing*, **8**, 601–612.
- [91] Lu, L. J., Xia, Y., Paccanaro, A., Yu, H., and Gerstein, M. (2005). Assessing the limits of genomic data integration for predicting protein networks. *Genome Research*, **15**(7), 945–953.
- [92] Luan, Y. and Li, H. (2003). Clustering of time-course gene expression data using a mixed-effects model with B-splines. *Bioinformatics*, **19**(4), 474–482.
- [93] Luscombe, N. M., Babu, M. M., Yu, H., Snyder, M., Teichmann, S. A., and Gerstein, M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, **431**, 308–12.
- [94] Ma, P., Castillo-Davis, C. I., Zhong, W., and Liu, J. S. (2006). A data-driven clustering method for time course gene expression data. *Nucleic Acids Research*, **34**(4), 1261–1269.
- [95] Mayoral, L. (2007). Minimum distance estimation of stationary and non-stationary ARFIMA processes. *The Econometrics Journal*, **10**(1), 124–148. February.
- [96] Mclachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley Series in Probability and Statistics. Wiley-Interscience.
- [97] Medvedovic, M., Yeung, K. Y., and Bumgarner, R. E. (2004). Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics*, **20**(8), 1222–1232.

- [98] M.J., B. and J.D., L. (2004). Chip-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, **83**(3), 349–360.
- [99] Myers, C., Barrett, D., Hibbs, M., Huttenhower, C., and Troyanskaya, O. (2006). Finding function: evaluation methods for functional genomic data. *BMC Genomics*, **7**(1), 187.
- [100] Nariai, N., Kim, S., Imoto, S., and Miyano, S. (2004). Using protein-protein interactions for refining gene networks estimated from microarray data by Bayesian networks. *Proceedings of the Pacific Symposium on Biocomputing*, pages 336–47.
- [101] Ng, S. K., Mclachlan, G. J., Wang, K., Jones, L. B.-T., and Ng, S.-W. (2006). A mixture model with random-effects components for clustering correlated gene-expression profiles. *Bioinformatics*, **22**(14), 1745–1752.
- [102] Opgen-Rhein, R. and Strimmer, K. (2007). Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process. *BMC Bioinformatics*, **8**(Suppl 2), S3.
- [103] Palmer, M. and Wu, Z. (1995). Verb semantics for English-Chinese translation. *Machine Translation*, **9**(4).
- [104] Parr, W. C. and Schucany, W. R. (1980). Minimum distance and robust estimation. *Journal of the American Statistical Association*, **75**(371), 616–624.
- [105] Parzen, E. (1962). On the estimation of a probability density function and mode. *Annals of Mathematical Statistics*, **33**, 1065–1076.

- [106] Price, C., Nasmyth, K., and Schuster, T. (1991). A general approach to the isolation of cell cycle-regulated genes in the budding yeast, *saccharomyces cerevisiae*. *Journal of Molecular Biology*, **218**(3), 543–556.
- [107] Qin, L. and Self, S. G. (2006). The clustering of regression models method with applications in gene expression data. *Biometrics*, **62**(2), 526–533.
- [108] Ramoni, M. F., Sebastiani, P., and Kohane, I. S. (2002). Cluster analysis of gene expression dynamics. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(14), 9121–9126.
- [109] Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, **11**, 95–130.
- [110] Rogers, D. F. and Adams, J. A. (1989). *Mathematical Elements for Computer Graphics*. McGraw-Hill Higher Education.
- [111] Rousseeuw, P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, **20**(1), 53–65.
- [112] Sain, S. (1994). Adaptive kernel density estimation.
- [113] Schäfer, J. and Strimmer, K. (2005). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, **21**(6), 754–64.
- [114] Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, **270**(5235), 467–470.

- [115] Schlicker, A., Domingues, F. S., Rahnenfuhrer, J., and Lengauer, T. (2006). A new measure for functional similarity of gene products based on gene ontology. *BMC Bioinformatics*, **7**, 302+.
- [116] Schliep, A., Costa, I. G., Steinhoff, C., and Schonhuth, A. (2005). Analyzing gene expression time-courses. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **2**(3), 179–193.
- [117] Scott, D. W. (2001). Parametric statistical modeling by minimum integrated square error. *Technometrics*, **43**(3), 274–285.
- [118] Scott, D. W. (2004). Outlier detection and clustering by partial mixture modeling. *COMPSTAT Symposium*, pages 453–465.
- [119] Segal, E., Barash, Y., Simon, I., Friedman, N., and Koller, D. (2002). From promoter sequence to expression: A probabilistic framework. pages 263–272. ACM Press.
- [120] Segal, E., Wang, H., and Koller, D. (2003a). Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics*, **19**(90001), 264i–272.
- [121] Segal, E., Shapira, M., Regev, A., Pe’er, D., Botstein, D., Koller, D., and Friedman, N. (2003b). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, **34**(2), 166–176.
- [122] Slonim, D. K. (2002). From patterns to pathways: gene expression data analysis comes of age. *Nature Genetics*, **32 Suppl**, 502–508.



- [123] Smith, P. (1982). *Curve fitting and modeling with splines using statistical variable selection techniques*. Report NASA 166034. NASA.
- [124] Smith, S., Fulton, D., Chia, T., Thorneycroft, D., Chapple, A., Dunstan, H., Hylton, C., and Smith, S. (2004). Diurnal changes in the transcriptom encoding enzymes of starch metabolism provide evidence for both transcriptional and posttranscriptional regulation of starch metabolism in arabidopsis leaves. *Plant Physiology*, **136**, 2687–2699.
- [125] Speer, N., Spieth, C., and Zell, A. (2005). Biological cluster validity indices based on the gene ontology. In *Advances in Intelligent Data Analysis*, pages 429–439.
- [126] Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology Cell*, **9**(12), 3273–97. Dec.
- [127] Sun, N., Carroll, R. J., and Zhao, H. (2006). Bayesian error analysis model for reconstructing transcriptional regulatory networks. *Proceedings of the National Academy of Sciences*, **103**(21), 7988–7993.
- [128] Tanay, A., Sharan, R., Kupiec, M., and Shamir, R. (2004). Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proceedings of the National Academy of Sciences*, **101**(9), 2981–2986.
- [129] Tang, N. and Vemuri, V. R. (2006). A web-knowledge-based clustering model

- for gene expression data analysis. *Proceedings of the Atlantic Web Intelligence Conference*.
- [130] Tarantola, A. (2006). Popper, Bayes and the inverse problem. *Nature Physics*, **2**(8).
- [131] Tavazoie, S., Hughes, J., Campbell, M., Cho, R., and Church, G. (1999). Systematic determination of genetic network architecture. *Nature Genetics*, **22**(3), 281–285.
- [132] The Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. *Nature Genetics*, **25**(1), 25–29.
- [133] Thalamuthu, A., Mukhopadhyay, I., Zheng, X., and Tseng, G. C. (2006). Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*, **22**(19), 2405–2412.
- [134] Tjaden, B. (2006). An approach for clustering gene expression data with error information. *BMC Bioinformatics*, **7**(17).
- [135] Troyanskaya, O. (2005). Putting microarrays in a context: Integrated analysis of diverse biological data. *Briefings in Bioinformatics*, **6**, 34–43.
- [136] Troyanskaya, O. G., Cantor, M., Sherlock, G., Brown, P. O., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**(6), 520–525.
- [137] Troyanskaya, O. G., Dolinski, K., Owen, A. B., Altman, R. B., and Botstein, D. (2003). A bayesian framework for combining heterogeneous data sources for

- gene function prediction (in *saccharomyces cerevisiae*). *Proceedings of the National Academy of Sciences*, **100**(14), 8348–8353.
- [138] Tseng, G. C. and Wong, W. H. (2005). Tight clustering: A resampling-based approach for identifying stable and tight patterns in data. *Biometrics*, **61**(1), 10–16. March.
- [139] Vogel, G. (2008). Breakthrough of the year: Reprogramming cells. *Science*, **322**(5909), 1766–1767.
- [140] von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S., Fields, S., and Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417**(6887), 399–403.
- [141] Wakefield, J., Zhou, C., and Self, G. (2003). Modelling gene expression data over time: Curve clustering with informative prior distributions. *Bayesian Statistics*, **7**, 721–732.
- [142] Walhout, A. J., Reboul, J., Shtanko, O., Bertin, N., Vaglio, P., Ge, H., Lee, H., Doucette-Stamm, L., Gunsalus, K. C., Schetter, A. J., Morton, D. G., Kempfues, K. J., Reinke, V., Kim, S. K., Piano, F., and Vidal, M. (2002). Integrating interactome, phenome, and transcriptome mapping data for the *c. elegans* germline. *Current Biology*, **12**(22), 1952–1958.
- [143] Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Monographs on Statistics and Applied Probability. Chapman and Hall, London.
- [144] Wang, H., Azuaje, F., Bodenreider, O., and Dopazo, J. (2004). Gene expression correlation and gene ontology-based similarity: an assessment of quantitative rela-

- tionships. *Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, pages 25–31.
- [145] Wang, J., Myklebost, O., and Hovig, E. (2003). Mgraph: graphical models for microarray data analysis. *Bioinformatics*, **19**(17), 2210–2211.
- [146] Werhli, A. V., Grzegorzcyk, M., and Husmeier, D. (2006). Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. *Bioinformatics*, **22**, 2523–2531.
- [147] Wichert, S., Fokianos, K., and Strimmer, K. (2004). Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics*, **20**(1), 5–20.
- [148] Wilkinson, D. J. J. (2007). Bayesian methods in Bioinformatics and computational systems biology. *Brief Bioinformatics*.
- [149] Wille, A., Zimmermann, P., Vranová, E., Fürholz, A., Laule, O., Bleuler, S., Hennig, L., Prelic, A., von Rohr, P., Thiele, L., Zitzler, E., Gruissem, W., and Bühlmann, P. (2004). Sparse graphical Gaussian modeling of the isoprenoid gene network in *Arabidopsis thaliana*. *Genome Biology*, **5**(11), R92.
- [150] Wilson, R. and Li, C.-T. (2003). A class of discrete multiresolution random fields and its application to image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **25**(1), 42–56.
- [151] Wu, F. X., Zhang, W. J., and Kusalik, A. J. (2005). Dynamic model-based clustering for time-course gene expression data. *Journal of Bioinformatics and Computational Biology*, **3**(4), 821–836.

- [152] Yeung, K., Fraley, C., Murua, A., Raftery, A., and Ruzzo, W. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics*, **17**(10), 977–987.
- [153] Yeung, K. Y., Medvedovic, M., and Bumgarner, R. E. (2003). Clustering gene expression data with repeated measurements. *Genome Biology*, **4**(5), R34.
- [154] Yu, J., Smith, V. A., Wang, P. P., Hartemink, A. J., and Jarvis, E. D. (2004). Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics*.
- [155] Yuan, Y. and Li, C.-T. (2007). Partial mixture model for tight clustering in exploratory gene expression analysis. *Proceedings of International Symposium on BioInformatics and BioEngineering*, pages 1061–1065.
- [156] Yuan, Y. and Li, C.-T. (2008a). A Bayes random fields approach for integrative large-scale regulatory network analysis. *Journal of Integrative Bioinformatics*, **5**(2), 99.
- [157] Yuan, Y. and Li, C.-T. (2008b). Partial mixture model for tight clustering of gene expression time-course. *BMC Bioinformatics*, **9**, 287.
- [158] Yuan, Y. and Li, C.-T. (2008c). Probabilistic framework for gene expression clustering validation based on gene ontology and graph theory. *Proceedings of International Conference of Acoustics, Speech, and Signal Processing*, pages 625–628.
- [159] Zacks, S. (1981). *Parametric Statistical Inference*. Pergamon Press.

- [160] Zellner, Arnold and Min, Chung-Ki (1995). Gibbs sampler convergence criteria. *Journal of the American Statistical Association*, **90**(431), 921–927.
- [161] Zhong, W. and Sternberg, P. (2007). Automated data integration for developmental biological research. *Development*, **134**, 3227–3238.
- [162] Zhu, Z., Pilpel, Y., and Church, G. M. (2002). Computational identification of transcription factor binding sites via a transcription-factor-centric clustering (tfcc) algorithm. *Journal of Molecular Biology*, **318**, 71–81.